

Prediction Model for Pipeline Pitting Corrosion Based on Multiple Feature Selection and Residual Correction

Zhenhao Zhu¹, Qiushuang Zheng², Hongbing Liu¹, Jingyang Zhang¹, Tong Wu¹ and Xianqiang Qu¹

Received: 02 May 2024 / Accepted: 08 August 2024

© Harbin Engineering University and Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

The transportation of oil and gas through pipelines is crucial for sustaining energy supply in industrial and civil sectors. However, the issue of pitting corrosion during pipeline operation poses an important threat to the structural integrity and safety of pipelines. This problem not only affects the longevity of pipelines but also has the potential to cause secondary disasters, such as oil and gas leaks, leading to environmental pollution and endangering public safety. Therefore, the development of a highly stable, accurate, and reliable model for predicting pipeline pitting corrosion is of paramount importance. In this study, a novel prediction model for pipeline pitting corrosion depth that integrates the sparrow search algorithm (SSA), regularized extreme learning machine (RELM), principal component analysis (PCA), and residual correction is proposed. Initially, RELM is utilized to forecast pipeline pitting corrosion depth, and SSA is employed for optimizing RELM's hyperparameters to enhance the model's predictive capabilities. Subsequently, the residuals of the SSA-RELM model are obtained by subtracting the prediction results of the model from actual measurements. Moreover, PCA is applied to reduce the dimensionality of the original 10 features, yielding 7 new features with enhanced information content. Finally, residuals are predicted by using the seven features obtained by PCA, and the prediction result is combined with the output of the SSA-RELM model to derive the predicted pipeline pitting corrosion depth by incorporating multiple feature selection and residual correction. Case study demonstrates that the proposed model reduces mean squared error, mean absolute percentage error, and mean absolute error by 66.80%, 42.71%, and 42.64%, respectively, compared with the SSA-RELM model. Research findings underscore the exceptional performance of the proposed integrated approach in predicting the depth of pipeline pitting corrosion.

Keywords Pipeline corrosion; Machine learning; Residual correction; Regularized extreme learning machine; Principal component analysis

1 Introduction

Pipeline transportation has become a primary method for long-distance energy transport, given its cost effectiveness, ease of management, and high level of safety (Arzaghi et al., 2018; Li et al., 2021; Peng et al., 2019). However, it is a double-edged sword because while it brings substantial economic benefits to local areas, it also poses potential threats to the surrounding environment and the safety of

residents (Al-Sabaei et al., 2023). According to the European Gas Pipeline Incident Data Group, nearly 27% of pipeline incidents are attributed to pipeline corrosion (Ma et al., 2023). Reports from the Pipeline and Hazardous Materials Safety Administration indicate that 20% of the 5 709 major pipeline incidents in the United States were due to corrosion issues (Peng et al., 2021). Consequently, pipeline corrosion has become an unavoidable issue in the comprehensive safety management of pipelines. Therefore, the establishment of a rational and effective prediction model for pipeline corrosion is crucial for reducing unnecessary pipeline excavations, mitigating the increased costs associated with frequent inspections, and accurately assessing the current safety status of pipelines (Ossai, 2020; Song et al., 2023).

The issue of pipeline corrosion generally encompasses internal and external corrosion. Insulation and cathodic protection layers are commonly applied externally to prevent external environmental corrosive substances from affecting pipelines. However, due to the complexity of the internal pipeline environment, which involves corrosive substances, the flow of oil and gas, and the presence of

Article Highlights

- We use PCA to extract the remaining information.
- Overfitting is prevented by applying RELM.
- In this study, the concept of residuals is incorporated into the field of pipeline pitting prediction, which effectively improves the accuracy of pipeline pitting depth prediction.

✉ Hongbing Liu
hb_liu@hrbeu.edu.cn

¹ Yantai Research Institute, Harbin Engineering University, Yantai 264005, China

² School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China

microorganisms, addressing internal corrosion is more challenging than addressing external corrosion (Jiang et al., 2017). Internal corrosion poses a greater risk to pipelines than external corrosion, prompting researchers to investigate methods for monitoring corrosion under complex conditions (May et al., 2022). Liu et al. (2018) employed high-frequency nondestructive testing techniques, such as piezoelectric pulse-echo and laser ultrasonics, and effectively identified the corrosion status of targets through the integration of time-frequency signal analysis algorithms. Tan et al. (2021) introduced distributed optical fiber sensors into pipelines to assess the presence of corrosion based on measured strain and locate areas of corrosion visually through strain distribution. Cadelano et al. (2016) utilized infrared thermography for pipeline corrosion monitoring, simulating real oil and gas pipeline conditions by analyzing the thermal effects of liquids within insulation layers to evaluate corrosion. While these experimental methods have enabled remarkable progress in corrosion monitoring, some pipelines are deeply buried underground or located in marine environments, making their employment challenging. Consequently, the utilization of relevant data to infer current corrosion status has become a preferred approach. The De Waard model for CO₂ corrosion is a classic corrosion empirical model proposed by De Waard. It determines the variation in steel corrosion rate in carbonic acid by considering factors such as temperature, CO₂ concentration, scaling, pH value, mass transfer, fluid flow rate, and flow effect (De Waard and Milliams, 1975). The American Society of Mechanical Engineers (ASME) developed the ASME B31G standard by referencing numerous cases. This standard calculates the remaining strength of pipelines on the basis of the size and characteristics of corrosion defects, thereby assessing the corrosion condition of pipelines (Mousavi and Moghaddam, 2020). The methods mentioned above primarily analyze the current corrosion situation by establishing linear models. However, these methods often exhibit certain limitations when facing the nonlinear problem of pipeline corrosion in reality (Ren et al., 2012).

In recent years, due to the rapid advancement of computer technology, a plethora of artificial intelligence models have been extensively employed in the field of pipeline corrosion to overcome the limitations of traditional models in addressing nonlinear complex problems (Li et al., 2022a; Zhang et al., 2022; Zheng et al., 2022). Mainstream artificial intelligence algorithms, such as Artificial Neural Networks (ANN), M5Tree, Bayesian Regularized Artificial Neural Networks, and Support Vector Machines (SVM), have been introduced into this domain (Ben Seghier et al., 2021; Chou et al., 2017; Li et al., 2022b). However, individual models often have certain limitations. Therefore, scholars have begun to combine artificial intelligence algorithms with other methods to obtain reasonable and effective results. Liu et al. (2021) compared three models: deci-

sion trees, ANN, and Bayesian networks. They found that incorporating probability distributions from Bayesian networks considerably enhanced the interpretability of models. Additionally, Li et al. (2021) selected the support vector regression algorithm as a small-sample prediction method and used the artificial bee colony algorithm for hyperparameter optimization to overcome the time-consuming nature of conventional grid search algorithms. Furthermore, they applied principal component analysis (PCA) to reduce redundancy among factors influencing corrosion, identifying the primary components with the greatest effect on underwater pipeline corrosion. Shi et al. (2017) based their work on four artificial intelligence algorithms, namely, multiple linear regression, ANN, random forest, and SVM, and utilized stacking ensemble to combine these models, establishing a pipeline performance prediction model based on machine learning and stacked ensemble modeling. This approach aimed to enhance the interpretability of the relationship between soil properties and pipeline performance and the predictive capability of the model.

Although the predictive capability of current artificial intelligence ensemble models has seen substantial progress, a certain residual error between predicted and actual values still exists. Therefore, many scholars have opted to use residual modeling for postprocessing (Liu and Chen, 2019). These postprocessing techniques primarily focus on handling residual sequences to derive residual correction values, which are then added to the prediction results of the previous combined model. The robustness of the results, considering residual correction, is remarkably enhanced through this approach (Xu et al., 2023a). Wei et al. (2020) used multiscale wavelet decomposition and residual sequence reconstruction, analyzed the high-frequency portion using a model based on SVM, and applied autoregressive integrated moving average for the low-frequency portion, thus constructing a combined model based on signal residual correction. Li et al. (2023) established a residual correction model based on Extreme Learning Machine (ELM), utilizing the residuals of partial training results on training sets to enhance the robustness of the ELM algorithm drastically, creating a residual-corrected ELM model.

In this study, a pipeline corrosion prediction model was developed on the basis of multiple feature selection and residual correction. The model utilized the SSA-RELM for the initial prediction of pipeline pitting corrosion depth, wherein the model's prediction residual was obtained by subtracting the actual measurement from the SSA-RELM prediction. Subsequently, PCA was employed to reduce the dimensionality of the original 10 features to derive 7 new features, which were used to predict the residual. Finally, the predicted residual was combined with the initial prediction results from SSA-RELM to obtain the prediction model for pipeline corrosion based on multiple feature selection and residual correction.

The remaining sections of this paper are as follows: A brief introduction to the methods used in this study is provided in Section 2. These methods include SSA, RELM, PCA, and residual correction. The evaluation criteria selected in this study are introduced in Section 3. A case study to demonstrate the feasibility and effectiveness of the proposed methods is presented in Section 4. The paper is concluded in Section 5.

2 Developed methodology

2.1 SSA

SSA is a heuristic optimization algorithm that simulates the searching and communication styles of sparrows. It solves problems by learning the behaviors of sparrows during foraging and group migration. Sparrows exhibit three types of foraging behaviors: 1) as explorers searching for food, 2) as followers joining explorers in the search for food, and 3) as sentinels deciding whether the population should continue foraging. Although explorers and followers have dynamic, interchangeable roles, their proportions remain constant. Explorers, as the leaders of the population's foraging, possess high fitness values and can explore a wide search area. Followers seek to obtain high fitness values by foraging alongside explorers, and some may continuously monitor explorers and seize food resources to improve their own predation rates. Generally, 10%–20% of the total number of sparrows in a population are established as sentinels, which alert the entire population of sparrows in time to avoid being preyed upon by predators after detecting danger (Xu et al., 2023b).

The update formula for the position of the discoverer during optimization iteration is

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \cdot \exp\left(-\frac{i}{\alpha \cdot \text{iter}_{\max}}\right) & \text{if } R_2 < \text{ST} \\ X_{i,j}^{t+1} + Q \cdot L & \text{if } R_2 \geq \text{ST} \end{cases} \quad (1)$$

where t represents the current iteration number; $\text{iter}_{\max}$ represents the maximum number of iterations; $X_{i,j}$ represents the position information of the i th sparrow in the j th dimension; $\alpha \in (0, 1]$ represents a random number; R_2 represents the warning value; ST represents the safety value; Q represents a random number following a normal distribution; and L represents a row vector with all elements being 1. When $R_2 < \text{ST}$, the discoverer searches extensively at this position for locations with good fitness values. When $R_2 \geq \text{ST}$, the discoverer updates the position randomly in accordance with a normal distribution.

The joiner position update formula is

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\frac{X_{\text{worst}}^t - X_{i,j}^t}{i^2}\right) & \text{if } i > P/2 \\ X_b^{t+1} + |X_{i,j}^t - X_b^{t+1}| \cdot A^+ \cdot L & \text{otherwise} \end{cases} \quad (2)$$

where X_b represents the current optimal position of the discoverer; X_{worst} represents the globally worst position; A represents a row vector with elements randomly assigned as 1 or -1 ; $A^+ = A^T(AA^T)^{-1}$. $i > P/2$, the joiner updates the position randomly in accordance with a normal distribution. Otherwise, the joiner moves near the current optimal position and participates in the search for locations with good fitness values.

The update description for the randomly selected sentinel's position is

$$X_{i,j}^{t+1} = \begin{cases} X_{\text{best}}^t + \beta \cdot |X_{i,j}^t - X_{\text{best}}^t| & \text{if } f_i > f_g \\ X_{i,j}^t + K \cdot \left(\frac{|X_{i,j}^t - X_{\text{worst}}^t|}{(f_i - f_w) + \varepsilon}\right) & \text{if } f_i = f_g \end{cases} \quad (3)$$

in the given formula, X_{best}^t represents the current global best position; β is the step-size control parameter and is a random number following a normal distribution with a mean of 0 and variance of 1; $K \in [-1, 1]$; f_i is the fitness value of the current sparrow individual; f_g and f_w are the current best and worst fitness values, respectively; and ε is a constant approaching 0. The sentinel moves from a position with low fitness toward the current best fitness position.

2.2 RELM

RELM is a further improvement on the ELM algorithm, which aims to suppress overfitting effectively by introducing a regularization term into the loss function. It has been widely used in engineering (He et al., 2019; Sun and Zhang, 2024; Wang et al., 2024). ELM is a typical single-hidden-layer feedforward neural network that is known for its fast training speed, eliminating the need to adjust connection weights, and excellent performance on high-latitude, large-scale datasets. It consists of three parts: the input, hidden, and output layers. Neural connections are present between the input and hidden layers, as well as between the hidden and output layers, as illustrated in Figure 1.

In the figure, $x_1, x_2, \dots, x_n$ denotes that the model selects a total of n feature outputs into ELM, and $y_1, y_2, \dots, y_m$ denotes that the model has a total of m output variables. h_l represents the neurons in the hidden layer, and l is the number of neurons in the hidden layer. w_{ij} and v_{jk} denote the connection weights between neurons.

The connection weight w between the input and hidden layers is presented in Equation (4):

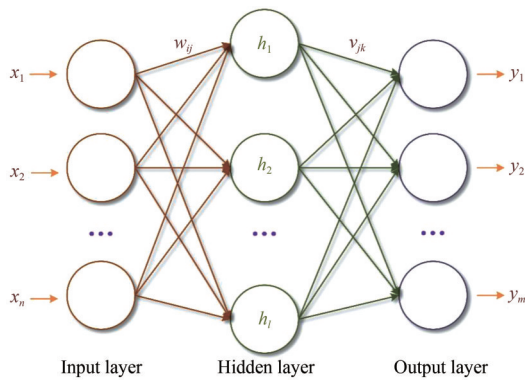


Figure 1 Diagram of ELM

$$\mathbf{w} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{l1} & w_{l2} & \cdots & w_{ln} \end{bmatrix} \quad (4)$$

The connection weight $\mathbf{v}$ between the hidden and output layers is presented as shown in Equation (5):

$$\mathbf{v} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & & \vdots \\ v_{l1} & v_{l2} & \cdots & v_{lm} \end{bmatrix} \quad (5)$$

The threshold $\boldsymbol{\eta}$ of the hidden layer neurons is presented as shown in Equation (6):

$$\boldsymbol{\eta} = [\eta_1 \quad \eta_2 \quad \cdots \quad \eta_l]^T \quad (6)$$

By combining the input and output values of the model to compose a matrix, the following matrices can be formed:

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1P} \\ e_{21} & e_{22} & \cdots & e_{2P} \\ \vdots & \vdots & & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nP} \end{bmatrix} \quad (7)$$

$$\mathbf{O} = \begin{bmatrix} o_{11} & o_{12} & \cdots & o_{1P} \\ o_{21} & o_{22} & \cdots & o_{2P} \\ \vdots & \vdots & & \vdots \\ o_{m1} & o_{m2} & \cdots & o_{mP} \end{bmatrix} \quad (8)$$

where P denotes the number of samples for the input model, and $\mathbf{E}$ and $\mathbf{O}$ are the input and output matrices, respectively.

Let $G(x)$ be the activation function of the hidden layer. The neural network output $\mathbf{T}$ can then be presented as

$$\mathbf{T} = [t_1 \quad t_2 \quad \cdots \quad t_P] \quad (9)$$

$$\mathbf{t}_j = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{nj} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^l v_{i1} G(\mathbf{w}_i \mathbf{e}_j + \eta_i) \\ \sum_{i=1}^l v_{i2} G(\mathbf{w}_i \mathbf{e}_j + \eta_i) \\ \vdots \\ \sum_{i=1}^l v_{im} G(\mathbf{w}_i \mathbf{e}_j + \eta_i) \end{bmatrix} \quad (10)$$

where $j = 1, 2, \dots, n$, $\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{in}]$, and $\mathbf{e}_j = [e_{1j}, e_{2j}, \dots, e_{nj}]$. The above formulas can be converted into the following expression:

$$\mathbf{M}\mathbf{v} = \mathbf{T}^T \quad (11)$$

In the above equation, $\mathbf{T}^T$ represents the transpose of matrix $\mathbf{T}$. $\mathbf{M}$ represents the output matrix of the hidden layer of the neural network, and its specific form is as follows:

$$\mathbf{M}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l, \eta_1, \eta_2, \dots, \eta_l, \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) = \begin{bmatrix} G(\mathbf{w}_1 \cdot \mathbf{e}_1 + \eta_1) G(\mathbf{w}_2 \cdot \mathbf{e}_1 + \eta_2) G(\mathbf{w}_l \cdot \mathbf{e}_1 + \eta_l) \\ G(\mathbf{w}_1 \cdot \mathbf{e}_2 + \eta_1) G(\mathbf{w}_2 \cdot \mathbf{e}_2 + \eta_2) G(\mathbf{w}_l \cdot \mathbf{e}_2 + \eta_l) \\ \vdots \\ G(\mathbf{w}_1 \cdot \mathbf{e}_n + \eta_1) G(\mathbf{w}_2 \cdot \mathbf{e}_n + \eta_2) G(\mathbf{w}_l \cdot \mathbf{e}_n + \eta_l) \end{bmatrix} \quad (12)$$

where $G(x)$ is the activation function of the hidden layer, $\mathbf{w}$ and $\boldsymbol{\eta}$ can be chosen randomly before training and then fixed during training, and $\mathbf{v}$ is obtained by using the least-squares solution (Huang et al., 2004).

When numerous outliers exist in training samples, the hidden layer output matrix may become ill-conditioned, which can affect the robustness and generalization of the model. Regularization theory can effectively address this issue. Therefore, introducing a regularization coefficient into the model can enhance its stability, yielding RELM (He et al., 2019).

2.3 PCA

PCA is a classic multivariate statistical method and an unsupervised dimensionality reduction technique. It achieves dimensionality reduction by linearly combining original features and eliminating redundant information, ensuring that the resulting composite features reflect the majority of

the information in the original features. The specific steps of this method under the assumption of a sample size of n with each sample having m -dimensional features are shown below (Ossai, 2019):

Step 1: The data in the sample are standardized to obtain the standardized result:

$$a_{ij} = (x_{ij} - \bar{x}_j) / h_j \quad (13)$$

where $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$, $\bar{x}_j$ is the sample means, and h_j is the standard deviation in the sample.

The standardized matrix A is constructed using a_{ij} , and its specific expression is

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad (14)$$

Step 2: The correlation matrix of the standardized matrix R can be obtained as

$$R = [r_{ij}]_{n \times m} = \frac{A^T A}{m - 1} \quad (15)$$

where $r_{ij} = \sum_{k=1}^m a_{kj} \cdot a_{ki} / (m - 1)$. The characteristic polynomial $|R - \lambda_j E| = 0$ is constructed to compute the eigenvalue λ_j .

Step 3: The cumulative contribution rate M can be obtained as

$$M = \sum_{j=1}^q \lambda_j / \sum_{i=1}^n \lambda_i \quad (16)$$

where M represents the cumulative amount of information extracted from the original features by the principal components, and q is the length of the reduced-dimensional data.

Step 4: The eigenvectors are computed.

2.4 Residual correction method

By incorporating factors with a considerable effect on pipeline corrosion into the machine learning model mentioned above, the predicted pipeline pitting corrosion, denoted as $y'_1(i)$, can be obtained. However, treating pipeline corrosion as a complex event and basing predictions solely on the relationship of corrosion effects with individual factors are insufficient. In practical scenarios, the combined effects of multiple factors and the influence of certain unknown factors must be considered. Capturing all potential influencing factors is challenging due to limitations in theoretical development and experimental conditions.

Therefore, this study employs a residual correction method, introducing the combined effects of multiple factors and utilizing PCA-reduced features as new predictive features. Through this approach, this research delves into predicting pipeline pitting corrosion, aiming to enhance predictive capabilities. The residual of the pipeline pitting corrosion is presented below (Yin and Liu, 2022):

$$r(i) = y_1(i) - y'_1(i) \quad (17)$$

where $r(i)$ represents the residual of the i th predicted value, whereas $y'_1(i)$ denotes the pitting corrosion of the i th sample pipeline.

By reintroducing $r(i)$ as the predicted value into the SSA-RELM algorithm for prediction, the model yields the residual prediction value $r'(i)$. Combining the residual prediction value $r'(i)$ with the initial predicted result $y'_1(i)$ produces the corrected predicted result $y_2(i)$, as shown in the following formula:

$$y_2(i) = y'_1(i) + r'(i) \quad (18)$$

2.5 Prediction model for pipeline corrosion based on multiple feature selection and residual correction

In accordance with the aforementioned theoretical framework, this investigation integrates four data-driven methodologies, namely, SSA, RELM, PCA, and residual correction, to construct a comprehensive model, as depicted in Figure 2. The specific procedure is delineated as follows:

Step 1: After an extensive review of established criteria and the integration of empirical data, this study ultimately identifies ten factors, such as redox potential (rp), pH value (ph), and pipe-to-soil potential (pp), as the model's features, with pipeline pitting corrosion depth as the output variable. The dataset is partitioned into training, validation, and testing sets with proportions of 60%, 20%, and 20%, respectively.

Step 2: RELM is employed as the predictive algorithm, and SSA is utilized for hyperparameter exploration.

Step 3: The residuals of the model are calculated by subtracting the predicted results of SSA-RELM from actual values.

Step 4: The original 10 features are subjected to dimensionality reduction to derive new features, which are subsequently utilized to predict the residuals obtained in Step 3.

Step 5: The predicted residuals from Step 4 are combined with the initial predictions from SSA-RELM in Step 2 to yield the final prediction. MAE, RMSE and MAPE are calculated. A comparative analysis is conducted between the results post residual correction and those obtained without residual correction to evaluate the performance of the model established in this study.

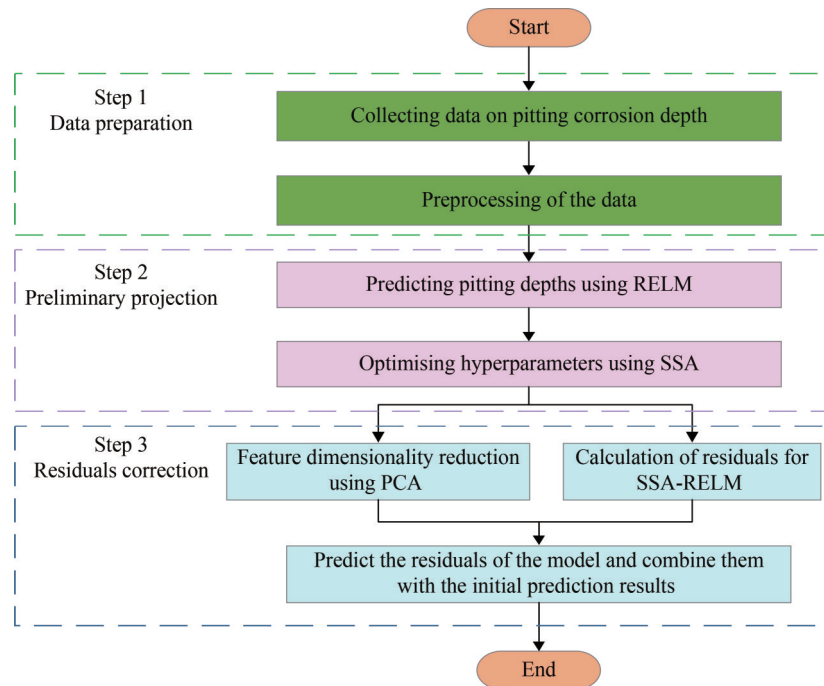


Figure 2 Framework of the proposed approach

3 Comparative criteria

Evaluation criteria that measure the model's generalization ability, known as performance metrics, are essential when evaluating the generalization performance of a model. These metrics can be used to assess the model's performance and serve as a basis for comparison between different models. This work selects MAE, RMSE, and MAPE as performance metrics.

$$\text{RMSE}(\delta, \hat{\delta}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\delta_i - \delta'_i)^2} \quad (19)$$

$$\text{MAPE}(\delta, \hat{\delta}) = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\delta_i - \delta'_i}{\delta_i} \right| \quad (20)$$

$$\text{MAE}(\delta, \hat{\delta}) = \frac{1}{n} \sum_{i=1}^n |\delta_i - \delta'_i| \quad (21)$$

In the formula, n represents the number of samples, δ_i denotes the actual value of the i th sample, and δ'_i represents the model's predicted value for the i th sample. Minimizing the RMSE, MAPE, and MAE of the model's pre-

dictions as much as possible is important to enhance the model's predictive performance.

4 Case study

4.1 Data collection

The establishment of a well-constructed and accurate database is crucial for enhancing the predictive capability of the prediction model for pipeline pitting corrosion depth. The publicly available database established by Velázquez has been widely utilized in the industry due to its comprehensiveness and rationality (Velázquez et al., 2010). This database encompasses the maximum pitting corrosion depth (pcd) of pipelines and 10 influencing factors, including redox potential (rp), pH value (ph), pipe-to-soil potential (pp), soil resistivity (re), water content (wc), bulk density (bd), dissolved chlorides (cc), bicarbonate concentration (bc), sulfate ion concentration (sc), and pipeline age (t). Table 1 provides a descriptive statistical summary of this database, wherein $X_{\max}$, $X_{\min}$, and X_{mean} represent the maximum, minimum, and mean values of variables, respectively.

Table 1 Descriptive statistics of the database

No.	t	ph	pp	re	wc	bd	cc	bc	sc	rp	pcd
$X_{\max}$	50.00	9.88	-0.42	399.50	66.00	1.56	672.70	195.20	1 370.20	348.00	13.44
$X_{\min}$	5.00	4.14	-1.97	1.90	8.80	1.10	1.00	1.00	1.00	2.10	0.41
X_{mean}	22.99	6.13	-0.88	50.15	23.90	1.30	47.73	19.67	152.97	167.04	2.02

The dataset comprises 259 samples. The data were randomly divided into the training, validation, and test sets in a 6:2:2 ratio, with 60%, 20%, and 20% of the data used for training, validation, and testing, respectively.

4.2 Comparison of model prediction

4.2.1 Comparison of the results of RELM with those of other algorithms

In this study, the RELM algorithm was employed for prediction, and SSA was utilized to determine the algorithm's hyperparameters. This approach replaced the traditional method of relying on the empirical determination of hyperparameters, thereby reducing the potential negative effect of unreasonable hyperparameter settings on the model's final prediction results. The maximum number of iterations for the SSA algorithm was set to 10 000, with 5 sparrows and upper and lower bounds for the parameter search range set at 1 000 and 0.001, respectively. The discovery rate was 0.2, and the warning value was 0.2.

The comparison of the prediction results of the SSA-RELM algorithm and other machine learning algorithms for the training set is shown in Figure 3. The graph shows that the SSA-XGBoost and SSA-ELM models exhibit results for the training set that closely align with the actual values. However, due to the complexity of these models, they perfectly adapt to the noise and details in the training data but overlook the overall trend of the real data, leading to overfitting and a decrease in generalization capability. Additionally, model overfitting can cause excessive sensitivity to outliers, thereby reducing the stability of the model. The RELM algorithm, which incorporates regularization, limits the model's complexity and reduces the risk of overfitting. This situation enhances the model's generalization capability and stability, resulting in superior performance on the test set.

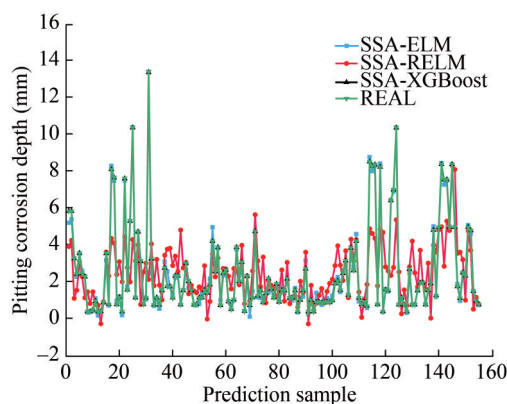


Figure 3 Comparison of the performances of each model on the training set

Table 2 presents a comparison of the performance metrics of several algorithms on the test set. It shows that the SSA-RELM model with regularization performs excep-

tionally well in terms of the MSE, MAPE, and MAE metrics. Specifically, compared with the other two models, SSA-RELM reduces MSE by 48.86% and 41.11%, respectively. Compared with those of SSA-ELM, the MAPE and MAE of SSA-RELM have reduced by 26.12% and 29.01%, respectively. These results indicate that the introduction of regularization has successfully mitigated overfitting issues, enhanced the model's generalization capability, improved prediction accuracy, and bolstered model stability, thereby effectively enhancing predictive capacity. These results are obtained because, with the introduction of regularization, both algorithms show a remarkable improvement in performance compared with the ELM algorithm without regularization. However, RELM offers a simpler model structure and lower computational complexity than XGBoost, which has a considerably higher model complexity, more complex tuning, and a consequent increase in exposure to a higher risk of overfitting. Therefore, RELM outperforms XGBoost when dealing with simple structured datasets, such as the prediction of pipeline pitting corrosion. However, despite SSA-RELM's outstanding performance in most aspects, the improvement in its MAPE and MAE is not as remarkable as that in its MSE because the logic of the optimization performed by SSA is to reduce MSE. The MAPE and MAE of SSA-RELM thus exhibit a slight increase compared with those of SSA-XGBoost (a difference of 30.57% and 8.87%). This situation implies that while SSA-RELM currently demonstrates good predictive performance, some issues that need further resolution remain.

Table 2 Comparison of model performance indices

Model	MSE	MAPE	MAE
SSA-ELM	2.628 1	1.314 6	1.296 6
SSA-XGBoost	2.282 4	0.743 8	0.845 4
SSA-RELM	1.344 1	0.971 2	0.920 4

4.2.2 Model prediction results based on residual correction

A prediction model for pipeline pitting corrosion depth was developed on the basis of SSA-RELM. While this model yields favorable prediction results, it still exhibits certain issues in terms of MAE and MAPE. Figure 4 shows the residuals obtained by using 60% of the training set to predict the outcomes on the 20% validation set and 20% test set. Figure 4 reveals that considerable residuals exist between the prediction results obtained through SSA-RELM and the actual data. This result is primarily attributed to the selection of 10 relatively independent influencing factors as features when constructing the data-driven prediction model for pitting corrosion depth in this study. However, the complex interactions among these factors and their comprehensive effect on pitting corrosion depth were not thoroughly explored. Apart from these 10 factors, some special unknown influencing factors that have not been

adequately considered may exist. This situation may lead to inaccuracies in the model under certain circumstances, necessitating further in-depth research and consideration of other potential factors.

This study employed a residual correction method to mitigate the effect of residuals and enhance the accuracy of predicting pipeline pitting corrosion depth. First, this study will delve into the internal relationships among the selected 10 influencing factors to derive new features with comprehensive information, enabling a thorough understanding of their effect on pitting corrosion depth. Subsequently, an analysis of these information-rich features will be conducted, providing the model with reasonable inputs to improve its fit with residuals, reduce residuals, and enhance predictive performance. Finally, the residual-corrected values will be combined with the original predicted results to obtain the residual-corrected prediction outcomes.

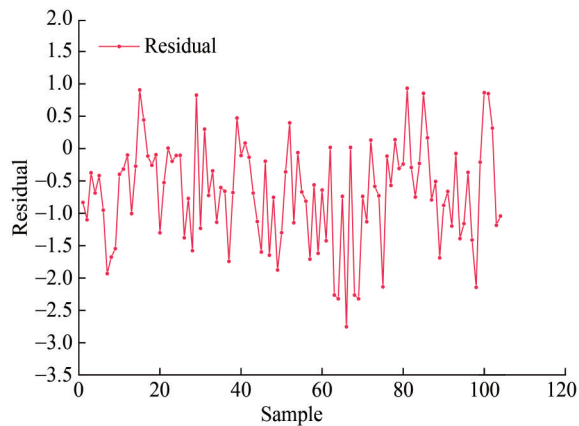


Figure 4 Residuals of SSA-RELM

In order to delve deeper into the composite effects of factors on pipeline pitting corrosion, this section employed PCA to reduce the dimensionality of the original 10 influencing factors. Figure 5 illustrates the cumulative contribution rate of features. In this study, a threshold of 90% cumulative contribution rate was utilized to select new features

after dimensionality reduction. As shown in Figure 5, 7 new features with a high contribution rate to pipeline pitting corrosion depth were chosen. These new features are all combinations of the original 10 influencing factors. By utilizing the original feature combinations to form new features, this study maximized the retention of information from the original features while reducing data dimensionality, to some extent considering the interactions among various influencing factors. This approach comprehensively captures the interrelated information among features, laying the groundwork for further enhancing the accuracy and robustness of the prediction model for pipeline pitting corrosion depth and thus adapting to the complexity of actual processes of pipeline pitting corrosion.

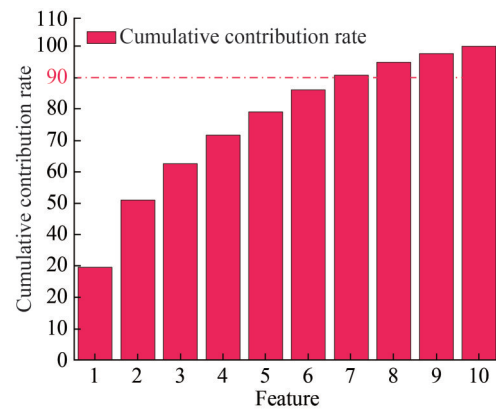


Figure 5 Cumulative contribution of features

The feature coefficients are presented in Table 3. Typically, the first principal component (Z1) explains the largest proportion of the variance, indicating that Z1 captures the most information in the data. Table 3 shows that the weight of *rp* in Z1 is considerably higher than those of other influencing factors. In practical engineering, *rp* directly affects the rate and direction of corrosion reaction; stability of electrode potential; formation of corrosion cells; and stability of the protective film, which has a considerable

Table 3 Feature coefficients of PCA

Factors	Z1	Z2	Z3	Z4	Z5	Z6	Z7
<i>t</i>	-0.270 60	-0.660 69	0.365 06	0.177 50	-0.419 67	0.150 18	-0.325 58
<i>ph</i>	-0.285 85	0.098 15	0.552 88	-0.368 62	-0.197 54	-0.141 18	0.633 40
<i>pp</i>	-0.233 63	-0.167 94	0.071 48	0.618 73	0.477 93	0.244 02	0.400 48
<i>re</i>	0.244 99	-0.069 54	-0.140 98	-0.130 66	-0.276 30	0.714 05	0.201 00
<i>wc</i>	-0.188 00	-0.122 61	0.036 21	0.034 67	0.128 09	-0.434 03	-0.206 68
<i>bd</i>	-0.006 52	0.662 32	0.442 26	0.345 19	-0.196 45	0.179 28	-0.342 17
<i>cc</i>	-0.147 90	-0.041 87	0.002 32	-0.341 09	0.363 16	0.215 49	-0.163 42
<i>bc</i>	-0.198 57	-0.012 05	0.179 14	-0.441 87	0.338 59	0.262 30	-0.318 89
<i>sc</i>	-0.156 63	0.092 83	0.237 35	0.026 43	0.331 12	0.203 06	-0.061 84
<i>rp</i>	0.780 54	-0.237 91	0.500 73	-0.007 76	0.262 55	-0.105 45	-0.001 27

effect on pipeline corrosion. This finding is consistent with the results of PCA. Additionally, the average weight of t from Z1 to Z7 is higher than that of other features. In practical engineering, pipelines gradually age as their service life increases, leading to issues such as the gradual deterioration of anticorrosion coatings or protective layers and the accumulation of internal deposits. Therefore, the increase in t is closely related to the extent of corrosion that pipelines experience.

The 20% validation set and 20% test set were combined with the residuals obtained earlier to create a new residual prediction dataset. By using PCA to reduce the dimensionality of the features in this dataset, seven new features are obtained and imported into the SSA-RELM model along with the residuals. Subsequently, the first 50% of the data were used for training to obtain new parameters, which were used for prediction along with the features included in the last 50% of the data to output the corresponding residuals. The residual prediction results are shown in Figure 6. The figure shows that the vast majority of the predicted residual values exhibit a high degree of proximity to the actual values. This outcome indicates that the application of PCA has effectively extracted crucial information from the original dataset, leading to the efficient and substantial dimensionality reduction of features. The resulting new feature set adequately encompasses the essential information required for residual prediction.

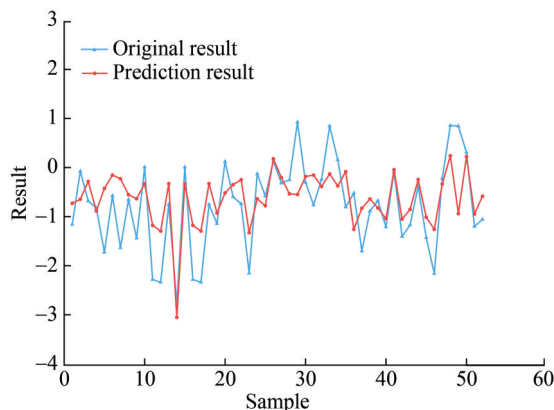


Figure 6 Residual prediction results

The corrected comprehensive prediction results were obtained by adding the residual prediction results to the predictions previously obtained by using the SSA-RELM algorithm. The comparison of the prediction results of the two algorithms is shown in Figure 7. The figure illustrates that the SSA-RELM model, without residual correction, exhibits outliers in some samples, highlighting the instability of the algorithm in predicting pitting corrosion. However, the algorithm with the inclusion of residual correction performs well on these samples, effectively avoiding the occurrence of outliers.

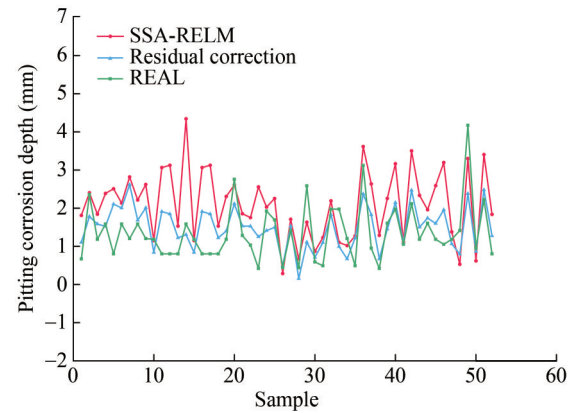


Figure 7 Comparison of the final prediction results

The performance comparison of the models is presented in Table 4. Compared with the SSA-RELM model alone, the model developed in this study demonstrates reductions of 66.80%, 42.71%, and 42.64% in MSE, MAPE, and MAE, respectively. These improvements in the series of performance metrics indicate that the SSA-RELM model, after residual correction, exhibits enhanced predictive accuracy, reduced overfitting, improved model stability, and decreased sensitivity to noise and outliers, thereby enhancing robustness. Additionally, the combination of PCA and residual correction well captures the nonlinear complex relationships in the data, providing reliable results for predicting pipeline corrosion depth and remarkably improving prediction accuracy.

Table 4 Comparison of performance indices between the original model and the model with residual correction added

Model	MSE	MAPE	MAE
SSA-RELM	1.344 1	0.971 2	0.920 4
Residual correction	0.446 2	0.556 4	0.527 9

5 Conclusions

This work combined the SSA-RELM-based prediction model for pipeline pitting corrosion with the PCA-based residual correction model to establish a prediction model for pipeline pitting corrosion based on multi-feature selection and residual correction. The reliability of this model was verified by using a real case of pitting corrosion data, leading to the following conclusions:

1) The findings demonstrate that the SSA-RELM-based pitting corrosion prediction model not only surpasses alternative machine learning algorithms in terms of MSE but also effectively mitigates overfitting concerns.

2) The concept of residuals was introduced into the prediction of pipeline pitting corrosion. Such an approach effectively improves the rationality of the model by incor-

porating information that is inadequately considered in the original model through residuals. It enables the model to consider various factors in complex environments comprehensively and thus predict pipeline pitting corrosion with increased accuracy. In practical engineering applications, this improvement helps cope with potential corrosion problems and reduce maintenance costs and risks.

3) The initial 10 independent features were amalgamated into seven new coupled features for predicting residuals through the utilization of PCA to reduce the dimensionality of the original feature set. This process effectively extracts the remaining information from residuals, allowing the model to capture key associations in the data accurately. In engineering applications, this data-processing method can reveal potential relationships between data well and improve the accuracy and reliability of prediction, thus providing strong support for decision-making.

4) The residual-corrected prediction model proposed in this research represents an efficacious and viable approach to the prediction of pipeline pitting corrosion, boasting minimal MSE, MAPE, and MAE of 0.446 2, 0.556 4, and 0.527 9, respectively. These values signify reductions of 66.80%, 42.71%, and 42.64% relative to the MSE, MAPE, and MAE of the conventional SSA-RELM algorithm. Additionally, the residual correction methodology yields a deepened understanding of the driving mechanisms behind pitting corrosion depth, culminating in precise and dependable practical predictions, thereby furnishing enhanced support for pipeline corrosion management.

Nomenclature

SSA	Sparrow Search Algorithm
RELM	Regularized Extreme Learning Machine
PCA	Principal Component Analysis
MSE	Mean Squared Error
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error

Funding Supported by the Natural Science Foundation of Shandong Province of China (ZR2022QE091), the Special fund for Taishan Industry Leading Talent Project (tsls20230605), Key R&D Program of Shandong Province, China (2023CXGC010407).

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

References

- Al-Sabaei AM, Alhussian H, Abdulkadir SJ, Jagadeesh A (2023) Prediction of oil and gas pipeline failures through machine learning approaches: A systematic review. *Energy Reports* 10: 1313-1338. <https://doi.org/10.1016/j.egy.2023.08.009>
- Arzaghi E, Abbassi R, Garaniya V, Binns J, Chin C, Khakzad N, Reniers G (2018) Developing a dynamic model for pitting and corrosion-fatigue damage of subsea pipelines. *Ocean Engineering* 150: 391-396. <https://doi.org/10.1016/j.oceaneng.2017.12.014>
- Ben Seghier MEA, Keshtegar B, Taleb-Berrouane M, Abbassi R, Trung NT (2021) Advanced intelligence frameworks for predicting maximum pitting corrosion depth in oil and gas pipelines. *Process Safety and Environmental Protection* 147: 818-833. <https://doi.org/10.1016/j.psep.2021.01.008>
- Cadelano G, Bortolin A, Ferrarini G, Molinas B, Giantin D, Zonta P, Bison P (2016) Corrosion detection in pipelines using infrared thermography: Experiments and data processing methods. *Journal of Nondestructive Evaluation* 35(3): 49. <https://doi.org/10.1007/s10921-016-0365-5>
- Chou JS, Ngo NT, Chong WK (2017) The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate. *Engineering Applications of Artificial Intelligence* 65: 471-483. <https://doi.org/10.1016/j.engappai.2016.09.008>
- De Waard C, Milliams DE (1975) Carbonic acid corrosion of steel. *Corrosion* 31(5): 177-181. <https://doi.org/10.5006/0010-9312-31.5.177>
- He C, Kang H, Yao T, Li X (2019) An effective classifier based on convolutional neural network and regularized extreme learning machine. *Mathematical biosciences and engineering* 16(6): 8309-8321. <https://doi.org/10.3934/MBE.2019420>
- Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: A new learning scheme of feedforward neural networks. 2004 IEEE international joint conference on neural networks, Budapest, Hungary, 985-990. <https://doi.org/10.1109/IJCNN.2004.1380068>
- Jiang T, Ren L, Jia ZG, Li DS, Li HN (2017) Pipeline internal corrosion monitoring based on distributed strain measurement technique. *Structural Control and Health Monitoring* 24(11): e2016. <https://doi.org/10.1002/stc.2016>
- Li X, Guo M, Zhang R, Chen G (2022a) A data-driven prediction model for maximum pitting corrosion depth of subsea oil pipelines using SSA-LSTM approach. *Ocean Engineering* 261: 112062. <https://doi.org/10.1016/j.oceaneng.2022.112062>
- Li X, Jia R, Zhang R, Yang S, Chen G (2022b) A KPCA-BRANN based data-driven approach to model corrosion degradation of subsea oil pipelines. *Reliability Engineering & System Safety* 219: 108231. <https://doi.org/10.1016/j.res.2021.108231>
- Li X, Zhang L, Khan F, Han Z (2021) A data-driven corrosion prediction model to support digitization of subsea operations. *Process Safety and Environmental Protection* 153: 413-421. <https://doi.org/10.1016/j.psep.2021.07.031>
- Li Z, Korovin I, Shi X, Gorbachev S, Gorbacheva N, Huang W, Cao J (2023) A data-driven rutting depth short-time prediction model with metaheuristic optimization for asphalt pavements based on RIOHTrack. *IEEE-CAA Journal of Automatica Sinica* 10(10): 1918-1932. <https://doi.org/10.1109/JAS.2023.123192>
- Liu G, Ayello F, Vera J, Eckert R, Bhat P (2021) An exploration on the machine learning approaches to determine the erosion rates for liquid hydrocarbon transmission pipelines towards safer and cleaner transportations. *Journal of Cleaner Production* 295: 126478. <https://doi.org/10.1016/j.jclepro.2021.126478>
- Liu H, Chen C (2019) Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy* 249: 392-408. <https://doi.org/10.1016/j.apenergy.2019.04.188>
- Liu H, Zhang L, Liu HF, Chen S, Wang S, Wong ZZ, Yao K (2018) High-frequency ultrasonic methods for determining corrosion layer thickness of hollow metallic components. *Ultrasonics* 89: 166-172. <https://doi.org/10.1016/j.ultras.2018.05.006>

- Ma H, Zhang W, Wang Y, Ai Y, Zheng W (2023) Advances in corrosion growth modeling for oil and gas pipelines: A review. *Process Safety and Environmental Protection* 171: 71–86. <https://doi.org/10.1016/j.psep.2022.12.054>
- May Z, Alam MK, Nayan NA (2022) Recent advances in nondestructive method and assessment of corrosion undercoating in carbon–steel pipelines. *Sensors* 22(17): 6654. <https://doi.org/10.3390/s22176654>
- Mousavi SS, Moghaddam AS (2020) Failure pressure estimation error for corroded pipeline using various revisions of ASME B31G. *Engineering Failure Analysis* 109: 104284. <https://doi.org/10.1016/j.engfailanal.2019.104284>
- Ossai CI (2019) A data-driven machine learning approach for corrosion risk assessment—A comparative study. *Big Data and Cognitive Computing* 3(2): 28. <https://doi.org/10.3390/bdcc3020028>
- Ossai CI (2020) Corrosion defect modelling of aged pipelines with a feed-forward multi-layer neural network for leak and burst failure estimation. *Engineering Failure Analysis* 110: 104397. <https://doi.org/10.1016/j.engfailanal.2020.104397>
- Peng S, Chen Q, Zheng C, Liu E (2019) Analysis of particle deposition in a new-type rectifying plate system during shale gas extraction. *Energy Science & Engineering* 8(3): 702–717. <https://doi.org/10.1002/ese3.543>
- Peng S, Zhang Z, Liu E, Liu W, Qiao W (2021) A new hybrid algorithm model for prediction of internal corrosion rate of multiphase pipeline. *Journal of Natural Gas Science and Engineering* 85: 103716. <https://doi.org/10.1016/j.jngse.2020.103716>
- Ren CY, Qiao W, Tian X (2012) Natural gas pipeline corrosion rate prediction model based on BP neural network. *Fuzzy Engineering and Operations Research*, Berlin Heidelberg, Germany, 449–455. https://doi.org/10.1007/978-3-642-28592-9_47
- Shi F, Liu Z, Li E (2017) Prediction of pipe performance with ensemble machine learning based approaches. 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control, Shanghai, China, 408–414. <https://doi.org/10.1109/SDPC.2017.84>
- Song Y, Wang Q, Zhang X, Dong L, Bai S, Zeng D, Xi Y (2023) Interpretable machine learning for maximum corrosion depth and influence factor analysis. *npj Materials Degradation* 7(1): 1–15. <https://doi.org/10.1038/s41529-023-00324-x>
- Sun Y, Zhang S (2024) A multiscale hybrid wind power prediction model based on least squares support vector regression–regularized extreme learning machine–multi-head attention–bidirectional gated recurrent unit and data decomposition. *Energies* 17(12): 2923. <https://doi.org/10.3390/en17122923>
- Tan X, Fan L, Huang Y, Bao Y (2021) Detection, visualization, quantification, and warning of pipe corrosion using distributed fiber optic sensors. *Automation in Construction* 132: 103953. <https://doi.org/10.1016/j.autcon.2021.103953>
- Velázquez JC, Caleyó F, Valor A, Hallen JM (2010) Field study—Pitting corrosion of underground pipelines related to local soil and pipe characteristics. *Corrosion* 66(1): 016001. <https://doi.org/10.5006/1.3318290>
- Wang J, Wang X, Zhang J, Shang X, Chen Y, Feng Y, Tian B (2024) Soil salinity inversion in Yellow River Delta by regularized extreme learning machine based on ICOA. *Remote Sensing* 16(9): 1565. <https://doi.org/10.3390/rs16091565>
- Wei B, Chen L, Li H, Yuan D, Wang G (2020) Optimized prediction model for concrete dam displacement based on signal residual amendment. *Applied Mathematical Modelling* 78: 20–36. <https://doi.org/10.1016/j.apm.2019.09.046>
- Xu B, Chen Z, Wang X, Bu J, Zhu Z, Zhang H, Lu J (2023a) Combined prediction model of concrete arch dam displacement based on cluster analysis considering signal residual correction. *Mechanical Systems and Signal Processing* 203: 110721. <https://doi.org/10.1016/j.ymssp.2023.110721>
- Xu B, Wang S, Xia H, Zhu Z, Chen X (2023b) A global sensitivity analysis method for safety influencing factors of RCC dams based on ISSA-ELM-Sobol. *Structures* 51: 288–302. <https://doi.org/10.1016/j.istruc.2023.03.027>
- Yin S, Liu H (2022) Wind power prediction based on outlier correction, ensemble reinforcement learning, and residual correction. *Energy* 250: 123857. <https://doi.org/10.1016/j.energy.2022.123857>
- Zhang M, Guo Y, Xie Q, Zhang Y, Wang D, Chen J (2022) Defect identification for oil and gas pipeline safety based on autonomous deep learning network. *Computer Communications* 195: 14–26. <https://doi.org/10.1016/j.comcom.2022.08.001>
- Zheng Q, Wang C, Liu W, Pang L (2022) Evaluation on development height of water-conducted fractures on overburden roof based on nonlinear algorithm. *Water* 14(23): 3853. <https://doi.org/10.3390/w14233853>