IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 September 2023, accepted 13 October 2023, date of publication 19 October 2023, date of current version 25 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326075

RESEARCH ARTICLE

Multi-Parameter Maximum Corrosion Depth Prediction Model for Buried Pipelines Based on GSCV-XGBoost

NIANNIAN WANG, LIUYANG SONG, HONGYUAN FANG, BIN LI, AND FUMING WANG

Yellow River Laboratory, Zhengzhou University, Zhengzhou 450001, China
National Local Joint Engineering Laboratory of Major Infrastructure Testing and Rehabilitation Technology, Zhengzhou 450001, China
Collaborative Innovation Center of Water Conservancy and Transportation Infrastructure Safety, Zhengzhou 450001, China

Corresponding author: Liuyang Song (sly_zzu@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3801000; in part by the National Natural Science Foundation of China under Grant 52108289 and Grant 51978630; in part by the Program for Innovative Research Team (in Science and Technology) in University of Henan Province under Grant 23IRTSTHN004; in part by the Program for Science and Technology Innovation Talents in Universities of Henan Province under Grant 23HASTIT006; in part by the Postdoctoral Science Foundation of China under Grant 2022TQ0306; in part by the Key Scientific Research Projects of Higher Education in Henan Province under Grant 21A560013; and in part by the Open Fund of Changjiang Institute of Survey, Lanning, Design and Research under Grant CX2020K10.

ABSTRACT Corrosion is one of the most common types of damage inburied oil and gas pipelines. Corrosion leaks can cause serious accidents and can be harmful to pipelinesduring service. The maximum corrosion depth of an oil and gas pipeline is an important indicator for assessing the remaining strength of the pipeline. An accurate prediction of the maximum corrosion depth is important for the safe operation of pipelines. Machine learning has been shown to perform well in predictive assessment efforts. However, previous studies have rarely considered the effects of corrosion characterization and parameter optimization simultaneously. In this study, a multi-parameter maximum corrosion depth prediction model for pipelines based on GSCV-XGBoost is proposed, which can be applied to real projects. The model performs feature extraction on the pipeline dataset through Pearson correlation analysis, identifies the parameters that contribute more to the maximum corrosion depth, and predicts the maximum corrosion depth of the pipeline using an optimized machine learning model. The machine learning model used in this study was obtained by optimizing the XGBoost model usingthe GirdSearchCV method. That is, the optimal hyperparameter combination of the model was obtained by 10-fold cross-validation and grid searching. The prediction results werecompared with those of five common machine learning models. The conclusions show that the GSCV-XGBoost model performs the best in predicting the maximum corrosion depth of the pipeline with the smallest error. The R^2 and Root Mean Square Error(RMSE)scores forthe test set were 0.9886 and 0.2057, respectively. The prediction accuracy was improved by 34.59% over that of the conventional XGBoost model.

INDEX TERMS Oil and gas pipelines, maximum corrosion depth prediction, machine learning, correlation analysis, XGBoost.

I. INTRODUCTION

Pipelines are an important transport medium in today's urban development and construction, and the safety of pipelines has always been the focus of attention from all walks of life. Oil and gas pipelines are important energy transmission infrastructure [1], [2], [3]. In the process of operation, oil and gas pipelines are very prone to destruction, such as corrosion and rupture [4], [5], [6] due to long-term internal pressure and the external environment. In recent years, oil and gas development and production conditions have become increasingly complex, and some oil and gas development technologies have resulted new corrosion problems. Pipeline corrosion and leakage problems occur frequently [7], [8], which seriously

The associate editor coordinating the review of this manuscript and approving it for publication was Frederico Guimarães.

impact pipeline safety and even the safety of people's lives and property.

The study of the residual strength of corroded pipelines helps assess the remaining life of the pipeline. This provides an important theoretical basis for the replacement of pipelines, which can effectively reduce or even prevent pipeline accidents. This is also important for improving the development of pipeline safety performance assessments [9]. This also facilitates the subsequent repair of damaged pipes in a timely manner for corrosion monitoring in various aspects. Much of the current research has focused on predicting the failure pressure of pipelines in order to assess their residual strength. Predicting the depth of corrosion of pipelines is also an important indicator for assessing the residual strength of pipelines, which helps determine the residual life of pipelines [10]. However, there are few studies in the literature that predict pipeline depth. The depth of corrosion in pipelines is commonly predicted using Artificial neural networks(ANN), SVM, Bayesian regression and Decision Tree. The integrated algorithm has a higher accuracy in handling predictive regression problems than traditional machine learning models [11]. Integration algorithms include three main categories: Bagging, Boosting, and Stacking. Integration algorithms have been used to improve machine learning by combining several models.

The motivation of this study is to consider the influence of various factors on the corrosion depth of pipelines and to realize an intelligent prediction of pipeline corrosion depth through an innovative machine learning model GSCV-XGBoost. This prediction model was obtained by optimizing the integrated learning algorithm, XGBoost, using the GirdSearchCV method. The remainder of this paper is organized as follows. Section III introduces the concepts and methods. Section V describes the study work. Section VI presents and discusses the results. Real application scenarios are presented in Section VII. Finally, Section VII concludes the study.

II. RELATED WORKS

Evaluating the residual strength of oil and gas pipelines with corrosion defects is essential for ensuring the safe operation and maintenance of pipelines. There is a long history of research on the residual strength of pipelines. Most research on assessing the residual strength of pipelines has focused on predicting the failure pressure of pipelines. Predicting the corrosion depth of a pipeline is another important method for assessing its residual strength. Machine learning is becoming increasingly important for these methods.

A. PREDICTING PIPELINE FAILURE PRESSURES

Predicting pipeline failure pressures helps determine the remaining strength and remaining life of a pipeline. Commonly used methods include experimental, finite element, and data-driven methods. Traditional standard codes and experimental methods for evaluating the residual strength of pipelines evaluation are sometimes conservative. The pipe failure pressure predicted by these methods is lower than the actual blast failure pressure, and premature pipe replacement can occur. This increases the cost of the project and does not provide good economic benefits in practical applications.

Some scholars have gradually introduced finite element ideas into the study and analysis of the residual value of pipelines and have achieved good results. Yang et al. [12] conducted a bending test study and finite element modeling of the static strength of corroded pipes, showing that the static strength of corroded pipes is significantly lower than that of normal pipes. The results of this study prove that the finite element method is significantly better than the traditional experimental method. Yang et al. [13] designed several different diameters and wall thicknesses of Q690 welded high-strength steel pipe model through the finite element modeling software ANSYS, proposed Q690 welded high-strength steel pipe cross-sectional longitudinal residual stress distribution model, the study of high-strength steel pipe residual working strength has important implications. Su et al. [14] studied the bursting capacity of a variety of steel tubes with corrosion defects under the action of internal pressure based on the finite element method by considering diameter, wall thickness and steel strength grade of the steel pipe. A failure pressure prediction method for medium and high strength corroded steel pipes was proposed. Since the rise of machine learning, it has been proven to perform well in various types of work [15], [16]. Kumar et al. [17] trained API 5L X80 pipeline dataset with different defect spacings,depths,defect lengths, and longitudinal compression loads using ANN. In addition, a failure pressure prediction model for highly ductile corroded pipes under combined loads was proposed. Lo et al. [18] used finite elements to simulate the failure of pipes with various corrosion geometrical parameters and loads, trained the finite element simulated pipe data using ANN and evaluated their performance. The evaluation results showed that this method has a low error. Shaik et al. [19] proposed an intelligent prediction model for the remaining life of crude oil pipelines using feedforward back propagation networks.The model can predict the conditions of crude oil pipelines based on specific factors, such as metal loss anomalies, wall thickness, weld anomalies and pressure flow rate, and calculate the remaining life of the pipeline based on the metal loss growth rate.

B. PREDICTING THE DEPTH OF CORROSION IN PIPELINES

The corrosion depth is an important index that affects the remaining strength of the pipeline, and it also has an important reference value for determining the remaining life of the pipeline. Moreover, corrosion depth has a greater impact on the pipeline than the corrosion length and corrosion width [20]. Fang et al. [21] showed that as the corrosion depth increases, the maximum principal stress and strain of the pipeline also increase, and the probability of pipe failure increases. Wang et al. [22] investigated the causes of cast iron pipe failure and established a new corrosion pit depth prediction model. It was demonstrated that with an increase in

corrosion depth, the toughness of the cast iron pipeline gradually decreases and the corrosion pit depth has more influence on the failure of pipeline than other factors. Ma et al. [23] established a new formula for predicting the damage pressure of corroded pipes made of high-strength steel materials, and found that the corrosion depth of pipes has a greater effect on the failure pressure of pipes than the corrosion width and length of pipes. Chen et al. [24] utilized the fractal hypothesis to establish a prediction model for the corrosion pit depth of nuclear power pipelines and laid down an accurate prediction model for the maximum corrosion pit depth of pipelines. This is of incredible significance in the study of the remaining work performance of corroded pipelines. The application of intelligent algorithms makes the study of pipeline corrosion depth more accurate and reliable [25]. Ma et al. [26] developed a PSO-SVM model to predict the growth of the corrosion depth in pipelines. Balekelayi et al. [27] researched the connection between the external corrosion depth of aging buried oil and gas pipelines and the outer soil factors. The expected model of the outer depth of the pipeline corrosion pit was determined using the Bayesian spectral analysis regression method. The prediction of the maximum corrosion depth is critical to the safety of defective pipelines, however, few studies have been conducted specifically on the maximum corrosion depth. Ben Seghier et al. [28] considered the effect of hyperparameter optimization on the model and developed a hybrid machine learning model SVR-FFA to predict the maximum pitting depth of pipelines. Velazquez et al. [29], [30] collected nearly 300 sets of data on the maximum pitting depth of pipelines with the corresponding pipeline characteristics and soil conditions, and analyzed the influence of local soil conditions and the characteristics of the pipeline itself on the pitting depth of the pipeline. An empirical prediction formula for the maximum pitting depth was established using multiple regression. Most models for predicting the corrosion depth of pipelines focus on conventional machine learning models such as ANN, SVM, and Bayesian regression, and do not simultaneously consider pipeline characteristics and hyper-parameter optimization.The XGBoost algorithm is an emerging and excellent integrated learning model that has a significant advantage in making various types of risk predictions [31], [32]. Compared with the research of Ben Seghier and Velazquez et al. the intelligent prediction model proposed in this study adopts a more advanced machine learning model and possesses better prediction capability. This study developed an intelligent prediction framework for the maximum corrosion depth of oil and gas pipelines based on this prediction model, which can be applied to a variety of practical engineering scenarios.

III. CONCEPTS AND METHODS

This section describes the intelligent prediction framework for the maximum corrosion depth of pipelines proposed in this study. The base model XGBoost applied to the intelligent framework is also presented, along with five other comparative machine learning models.

A. MAXIMUM CORROSION DEPTH PREDICTION FRAMEWORK

Many factors affect the maximum corrosion depth of the pipe, including the pipe and soil characteristics, involving multiple variables related to the maximum corrosion depth. The use of too many redundant variables affects the accuracy of the subsequent model predictions and requires feature engineering. This includes a correlation analysis of variables and feature extraction [33] to extract important variables for the next step of model building. The advanced XGBoost algorithm in the field of machine learning is used to build the prediction model and divide the training set and test set. Model parameter tuning of the training set is an essential step in regression prediction, and parameter optimization can improve the predictive power of the model [34]. Subsequently, the optimized GSCV-XGBoost model was used to predict the maximum corrosion depth test set of the pipe. Finally, an error analysis is performed on the prediction results to compare the difference between the predicted value of the model and the true value of the sample to evaluate the prediction ability of the model. In this study, an intelligent prediction model of the maximum corrosion depth of a pipeline is proposed for the pipeline corrosion problem, as shown in Figure 1. It includes five major parts: data collection, feature engineering, machine learning modeling, parameter optimization and model training, and evaluation of prediction results.

- Data collection. The data collection process is the foundation of machine learning modeling, and the model can be trained and predicted only after a reasonable data-set is input into the model. The data source used in this study was pipe and soil sample data collected by researchers during a three-year period in the field excavation of an onshore buried pipeline in southern Mexico. The maximum depth of corrosion loss for each pipeline with a diameter equal to or less than twice the wall thickness was measured and collected, along with the corresponding pipeline characteristic variables, including pipeline design, operational, and environmental data.
- Feature engineering. Through a correlation analysis of the maximum corrosion depth of the pipe and the related feature variables, some feature variables with strong correlation with the maximum corrosion depth were extracted and used as the input variables of the model.
- Machine learning modeling. After feature extraction, the dataset is divided into a training set and a test set, and the XGBoost algorithm, which is currently emerging in the field of machine learning, is used to make preliminary modeling predictions of the pipe corrosion depth in the dataset.
- Parameter optimization and model training.Machine learning models often have many hyperparameters, and the value of the hyperparameters will also significantly affect the prediction results [35]. The hyperparameter optimization of the model and selection of the optimal combination of hyperparameters for model training can improve the prediction accuracy of the model.

- Prediction results evaluation.The optimized GSCV-XGBoost model was used to predict the maximum corrosion depth of pipes on the test set. The results of the prediction results are analyzed and various error indicators are calculated. The error metric is used to measure the difference between the predicted results of each model and the true [36], [37], which is the difference between the predicted output value and the true value of the sample. The prediction accuracy and error of the model on the test set were comprehensively using several different error evaluation indicators.

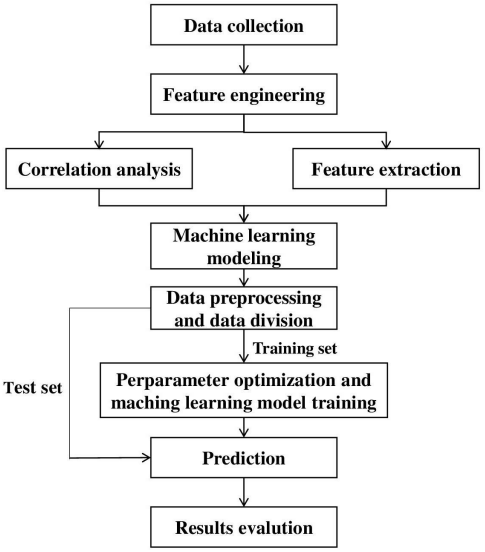

FIGURE 1. Intelligent prediction model of maximum corrosion depth of the pipeline.

B. XGBOOST - BASE MODEL FOR MAXIMUM CORROSION DEPTH PREDICTION

XGBoost, whose full name is eXtreme Gradient Boosting, is an efficient boosting integrated learning [38], [39], [40]. The core idea of the boosting algorithm is to aggregate multiple weak learners to form a strong learner, so that the strong learner can obtain the advantages of various weak learners to achieve optimal model performance. XGBoost is essentially a boosted tree model built on the GBDT model that update the model by minimizing losses. Compared to the traditional GBDT model, XGBoost has made significant improvements in algorithm accuracy, speed and generalization ability. XGBoost can efficiently process billions of data in a parallel and distributed fashion; it can also has handle sparse data and parallel learning using column blocks [41]. In this model, XGBoost extends the loss function to a second-order derivative, making the model closer to its true loss. Moreover, the loss function adds a regularization term to the basic empirical loss to optimize the complexity of the model and prevent overfitting. XGBoost can also select the optimal features utilizing information gain, and make the objective function continuously approach the optimal value by analyzing the learning error of the base model CART regression tree and constantly updating the sample weights at each iteration [42]. The objective function is continuously approximated by analyzing the learning error of the base model CART regression tree and updating the sample weights at each iteration. Compared with other conventional machine learning methods, XGBoost has the unique advantage of significantly enhancing the generalization ability of the model and improving the accuracy of the model prediction. XGBoost has many hyperparameters, and parameter optimization is extremely important for improving the accuracy of the model. The model participates in the adjustment of various hyperparameters such as the number of basic tree models (the maximum number of iterations), maximum depth of the tree, and gamma.

This study is a regression problem, where the maximum corrosion depth of the pipe is y, and the important variables related to y are $X^1, X^2, X^3 \cdots X^N$, then the relationship between the input and output of this model is:

This study is a regression problem, the maximum corrosion depth of the pipe is y, and the important variables related to y are $X^1, X^2, X^3 \cdots X^N$. Thethe relationship between the input and output of this model is:

$$f\left(X^1, X^2, X^3 \cdots X^N\right) = y \tag{1}$$

The input vector set of this prediction model $D = \{(x_i, y_i)\}$, where each weak learner is denoted by fk, and the maximum predicted corrosion depth of the ith sample obtained after the superposition of K weak learners is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{2}$$

Suppose the base model for the t-iteration is:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

The basic form of the XGBoost loss function consists of an empirical loss term and a regularization term (the sum of the complexitiesof decision trees):

$$L = \sum_{i=1}^{n} I(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{4}$$

Then take step t as an example, the loss function can be rewritten as:

$$L^{(t)} = \sum_{i=1}^{n} I(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{Constant} \tag{5}$$

A second-order Taylor expansion of the loss function for the first half, which is also an important feature of XGBoost, can be rewritten as:

$$1\left(y_i, y_i^{(\hat{t-1})} + f_t(x_i)\right) 1\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \tag{6}$$

where g_i represents the first-order derivative of the loss function, and h_i represents the second-order derivative of the loss function. Here we have used the squared loss function using:

$$l\left(y_i, \hat{y}_i^{(t-1)}\right) = \left(y_i - \hat{y}_i^{(t-1)}\right)^2 \tag{7}$$

Substituting the second-order Taylor expansion into the loss function of sub-equation (4):

$$L^{(t)} \approx \sum_{i=1}^{n} \left[I(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) \right] + \Omega(f_t) + \text{Constant} \tag{8}$$

The loss function is simplified by removing the constant term as:

$$L^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{9}$$

It is also necessary to continue to derive the regression decision tree node splitting condition, assuming that each CART regression tree contains the weights w of its leaf nodes and a kind of sample-to-leaf node mapping relationship q. Here the mapping relationship can be expressed as the branching structure of the decision tree, and the model complexity Ω can be expressed in terms of the number of leaf nodes T with weights w, then we have:

$$f_t(x) = w_{q(x)} \tag{10}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{11}$$

where $j = 1, 2, 3 \ldots T$, the jth leaf node of the t-tree regression decision tree contains the following set of samples:

$$I_j = \{i | q(x_i) = j\} \tag{12}$$

Define $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$, where G_j and H_j denote the cumulative first-order partial derivative and cumulative second-order partial derivative values of the samples in the jth leaf node, respectively, the loss function can be written as:

$$L^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma \lambda \tag{13}$$

We take out the leaf node j separately at this point: $G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2$, when the leaf nodes of each tree independent of each other reach the optimal value, the loss function of the whole model will also reach the optimal value. It is derived and the result equals 0, giving the optimal point w_j^* and the optimal value L.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{14}$$

$$L = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma \lambda \tag{15}$$

Suppose that the regression decision tree undergoes a feature split at a leaf node, and the loss function before its split is expressed as:

$$L_{\text{before}} = -\frac{1}{2} \left[\frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] + \lambda \tag{16}$$

The loss function after splitting is given by:

$$L_{\text{after}} = -\frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} \right] + 2\lambda \tag{17}$$

Then the information gained after splitting can be expressed as:

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \lambda \tag{18}$$

If the information gain $\text{Gain} > 0$, then it means that the objective function value becomes smaller after splitting into two leaf nodes, andthe result of this split is considered, otherwise it is not considered. It is necessary to iterate all features to determine the optimal splitting feature.

C. MODELS FOR COMPARISON

When evaluating the prediction results of a model, the error parameter metrics of the model are calculated. It is also necessary to include different machine learning regression models for comparison. Five commonly used regression prediction models, BP Neural Network, Support Vector Regression, Decision Regression Tree, Random Forest and AdaBoost, are chosen for comparison models.

1) BP NEURAL NETWORK (BP)

BP neural network is a multilayer feedforward neural network by the reverse transmission of errors [43]. The key steps include forward propagation of the signal and reverse transmission of error, where reverse transmission of error is the core step of neural network training. The input values of the neural network are computed and passed into the hidden layer and output layer in turn, and the error between the final output value and the desired output value is returned along the original route.And the error is minimized by modifying the threshold of the neurons and the weights of the neuron connections in each layer [44]. In this study, the regression prediction model of a BP neural network is established for the maximum depth data set of pipe corrosion.

2) SUPPORT VECTOR REGRESSION (SVR)

The SVM model is a widely used dichotomous model that can address classification and regression problems [45]. Support vector regression machine SVR is an important branch of support vector machines that specialize in solving regression problems. SVR creates an ''interval band'' on both sides of the linear function, and the loss is not calculated for the data within the interval band. To maximize the distance to the farthest point of the hyperplane, the optimized model is finally obtained by minimizing the total loss and maximizing

TABLE 1. Model error analysis indicators.

Statistical metrics	Notation	Expression
Explained Variance	EV	$1-\frac{\sum_{i=1}^{n}((y_i-\hat{y}_i)-(\bar{y}-\bar{\hat{y}}))^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
Coefficient of Determination	R^2	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}$
Mean Square Error	MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
Mean Absolute Error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$
Median Absolute Error	MedAE	$\text{median}(\lvert y_i-\hat{y}_i\rvert)$

Note: n denotes the total number of samples; p denotes the total number of features; y_i denotes the true value of the sample; $\hat{y}_i$ denotes the model prediction value; $\bar{y}_i$ denotes the arithmetic mean of the true value; $\bar{\hat{y}}$ denotes the arithmetic mean of the predicted value.

the interval. Important hyperparameters of the SVR model include C and epsilon.

3) DECISION TREE REGRESSION (DTR)

Decision Tree [46] is a common supervised learning method. Data record sets are organized into a hierarchical structure consisting of a series of rule-governed nodes and branches [47]. A complete decision tree comprises nodes and directed edges, with the root node comprising a complete set of samples, leaf nodes representing categories, and internal nodes representing features. Starting from the root node, a particular feature is selected to distribute the samples. Further splitting is performed where each child node acts as a parent node from which more nodes and decision tree layers are generated until the classification of all samples is complete [48]. Decision Tree Regression used to solve the regression problem. Important hyperparameters for the DTR model include the depth of the tree, the minimum number of samples required for segmentation.

4) RANDOM FOREST (RF)

Random Forest is based on bootstrap sampling and is an extended variant of bagging. This efficient integrated learning algorithm can be used for multi-class classification and regression.The advantages of the RF method are: the principle of the algorithm is relatively simple and easy to implement. Compared with BP and SVR, it is suitable for more feature parameters and has better performance in practical applications [49]. According to the characteristics, the data samples are continuously divided according to certain conditions, and finally the purpose of classification or regression is achieved. The RF algorithm obtains different subsets of samples by randomly sampling the original pipeline maximum corrosion depth dataset, training the base learner for each subset of samples and performing model integration and outputs. The final output is obtained by voting to solve the classification problem. And the final model output is obtained by averaging the output of the base learners when solving the regression problem, which greatly improves the model performance of random forest greatly improved compared to that of individual learners [50]. In this model, the random forest uses a regression decision tree as its base learner to fit the regression problem. Important hyperparameters of the RF model are the depth of the tree and the number of learners.

5) ADABOOST

AdaBoost [51] is a typical representation of the boosting integration algorithm. AdaBoost completes two important steps when working, one is to increase the weight of the sample that was classified incorrectly by the weak classifier in the previous round, and to reduce the weight of the correctly classified sample.The second is to linearly combine multiple weak classifiers to increase the weight of weak classifiers with a good classification effect and to reduce the weights of weak classifiers with high classification error rate [52]. In this model, AdaBoost also uses decision trees as the basis for its weak learners, which can fit the regression problem well with its unique advantages. Common hyperparameters of the AdaBoost model include the learning rate and number of weak learners.

D. ERROR ANALYSIS INDICATORS

The results of machine learning predictions must tested using error analysis. In this study, the deviation between the model prediction results and true values is measured using several error analysis parameter indicators. The error parameter metrics used in this study are Explainable Variance Value (EV), Coefficient of Determination (R^2), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Median Absolute Error (MedAE). Table 1 lists the specific expressions for these error metrics. EV and R^2 denote the accuracy of fitting the model sample values

TABLE 2. Distribution of parameters.

	t (y)	ph	wc (%)	sbd (g/ml)	dcc (ppm)	rp (mV)	psp (V)	sr (Ωm)	bc (ppm)	sc (ppm)	mpd (mm)
Min	5	4.14	8.80	1.1	3.2	2.1	0.42	1.9	3.91	2.57	0.41
25%	19	5.42	19.40	1.24	11.45	99.5	0.74	15.3	8.32	60.23	0.81
50%	21	6.08	23.5	1.31	19.02	155	0.86	31.6	11.1	109.3	1.24
75%	26	6.81	27.7	1.37	43.21	235	1.01	64.25	18.48	181.7	2.16
Max	50	9.88	66	1.56	672.7	348	1.97	399.5	195.2	1370.2	13.44
Avg	22.88	6.14	23.9	1.3	47.78	167.13	0.88	50.24	19.92	155.25	1.98
Std	9.12	0.93	6.70	0.09	75.55	85.81	0.24	56.11	25.44	168.49	1.98
Ske	0.75	0.52	1.51	0.11	3.99	0.23	1.18	2.54	3.74	3.34	2.56
Kur	0.64	0.55	7.36	-0.19	22.29	-0.88	4.09	8.73	17.25	15.70	7.51

TABLE 3. The value range of |PCC| and the corresponding degree of correlation.

|PCC|	[0,0.2]	[0.2,0.4]	[0.4,0.6]	[0.6,0.8]	[0.8,1.0]
Degree of correlation	Very weak or no correlation	Weak correlation	Medium correlation	Strong correlation	Very strong correlation

to the predicted results, and the larger the value calculated according to the formula, the more accurate the model is. MSE, RMSE, MAE, and MedAE denote the prediction error of the model, the smaller the value, the smaller the error. MSE measures the average of the squared difference between the predicted value and the true value, which is more sensitive to the samples with larger errors.RMSE is the square root of MSE, which is in the same units as the original data, is easier to interpret, and is more commonly used. MAE is used to measure the magnitude of the average model error. MedAE is used to measure the median of the absolute values of errors for all samples.

IV. EMPIRICAL ANALYSIS

A. EXPERIMENTAL DATA COLLECTION

All data in this study were obtained from the literature [29], [30], the researchers Velazquez JC et al. collected 259 sets of soil and pipeline samples over a three-year period, which were from onshore buried oil and gas transportation pipelines in service in southern Mexico. The soil samples included six types of clay, clay loam, sandy clay loam, chalky clay, powdered clay loam, and silt loam, and the pipeline samples had the maximum corrosion depth and age. In addition to the four groups of experimental missing data, the remaining 255 groups of pipeline characteristic data were used as the object of study in this study. The parameters related to the pipeline include the following:

Maximum pitting depth (mpd) and pipe exposure time (t); parameters related to soil properties include: PH (ph), redox potential (rp), pipe-to-soil potential (psp), soil resistivity (sr), water content(wc), soil bulk density(sbd), dispersed chloride content(dcc), bicarbonate content (bc), sulfate content (sc), and coating type (ct). Among them, the coating type (ct) is different from the other parameters, which are sub-typed variables and it is inconvenient to count their distribution. The data of the pipe-to-soil potential (psp) are all negative values, and for the convenience of calculation, this study calculates all the data of psp by taking their opposite numbers. The distributions of the remaining parameters are presented in Table 2. Avg stands for mean, Std is the standard deviation, Ske is the skewness, and Kur is the kurtosis. The histogram of each parameter is shown in Figure 2. The scatter plot of each parameter with the maximum pitting depth is shown in Figure 3.

B. FEATURE ENGINEERING

If the training set data aretoo complex, machine learning is prone to overfitting, which requires dimensionality reduction of the data to extract important variables and reduce the complexity of the model. Feature selection [53], [54] is an important data processing method, and filtering is the most commonly used feature selection method forremovingirrelevant and redundant variables, reducingthe dimensionality of the data, reducingthe running time of the model andimproveingthe generalization ability of the model. Pearson correlation analysis isthe filtered approach, which is easy to calculate and can effectively measure the relationship between features and response variables.

To further understand the relationship between the data, clarify the relationship between the maximum corrosion depth and each of the remaining parameters, and identify the input variables for the next step of machine learning, Pearson correlation analysis needs to be performed on each data variable [55]. The strength of the correlation between variables can be judged by calculating the Pearson correlation coefficient, and the calculation formula refers to Equation 19 and Equation 20. The larger the absolute value of the calculated Pearson correlation coefficient, the stronger the correlation between the variables, and conversely, the lower the correlation. The closer the correlation coefficient is to 0, the lower the correlation. The closer the correlation coefficient is to 1 and -1, the stronger the correlation. A positive correlation coefficient indicates that the relationship between variables is positively correlated, and a negative correlation coefficient indicates that there is a negative correlation

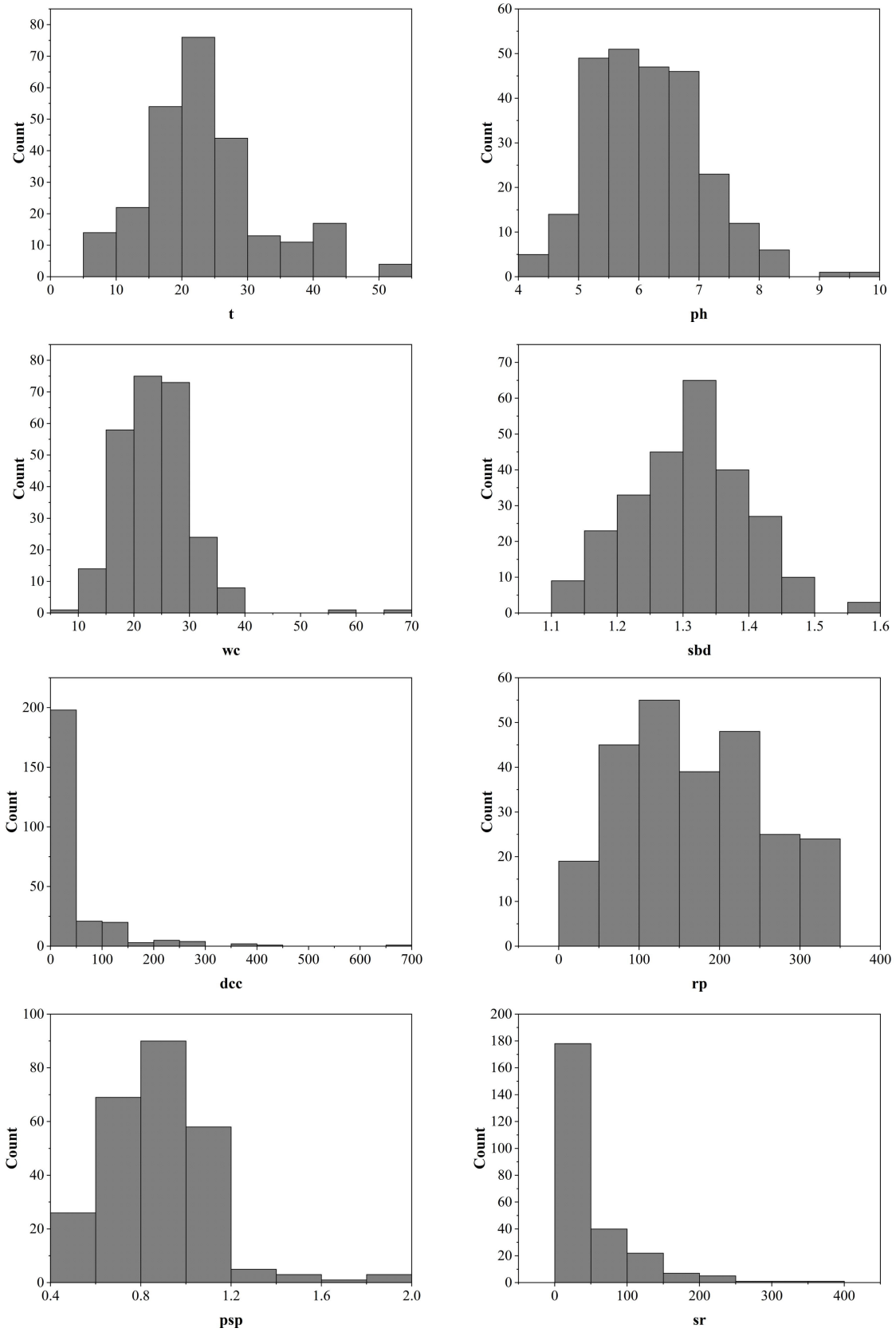

FIGURE 2. Histogram of each parameter.

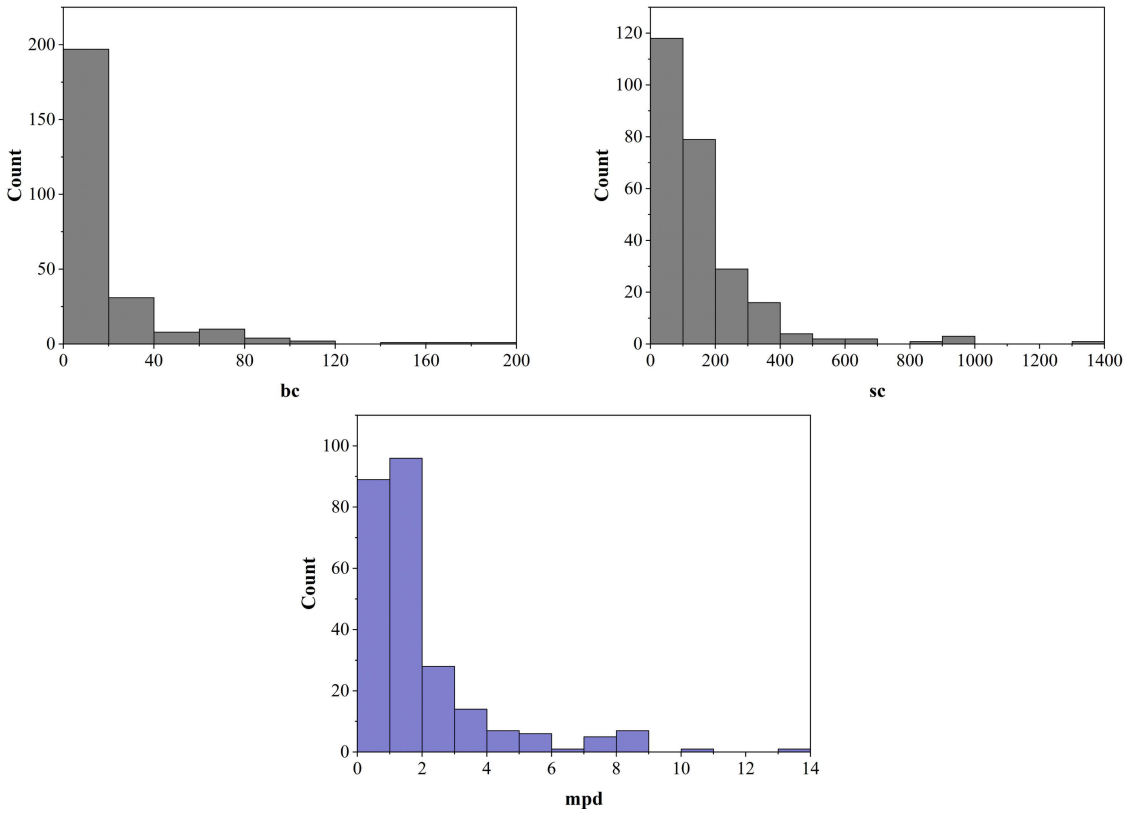

FIGURE 2. *(Continued.)* **Histogram of each parameter.**

between variables. The correlation coefficient is expressed in PCC, and the absolute value of the correlation coefficient is expressed in |PCC|. Table 3 shows the value range of |PCC| and the corresponding degree of correlation. Correlation analysis was performed and plotted by Python, as shown in Figure4.

Pearson correlation coefficient(PCC):

$$\mathrm{PCC}(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{cov}(\mathrm{X},\mathrm{Y})}{\sigma_x \sigma_y} \tag{19}$$

$$\mathrm{cov}(\mathrm{X},\mathrm{Y}) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n} \tag{20}$$

where cov(X, Y) is the covariance of the variables X, Y, and $\overline{X}$ is the mean of X, and $\overline{Y}$ is the mean of Y, and σ_x is the standard deviation of X, and σ_y is the standard deviation of Y.

As shown in Figure 4, the correlation between the variables and the maximum corrosion depth (mpd) was ranked from strong to weak: t > sbd > wc > dcc > ph = psp > rp > sr > bc > sc. The correlation coefficients of the three variables sr, bc, and sc are close to zerofor mpd, therefore the three parameters sr, bc, and sc need to be removed. In Figure 3, it can also be observed that the correlation between each parameter of sr, bc,sc and mpd is particularly weak based on the scatter plot of each variable versus the maximum corrosion depth of the pipeline. The remaining sevenparameters (t, sbd, wc,dcc,ph, pspand rp) that contribute more to the maximum corrosion depth wereselected as the control variables for mpd and the input variables for the model. As shown in Figure 4, the exposure time of the pipeline(t), soil bulk density(sbd), and water content(wc) significantly influence on the maximum corrosion depth of the pipeline.

C. MODEL PARAMETER OPTIMIZATION

In actual machine learning model training, to improve the generalization ability of the model and reduce the generalization error of the model as much as possible, it is necessary to solve the problem of hyperparameter optimization. Different hyperparameters often yield different results, and choosethe optimal combination of models. In this study, we choose GirdSearchCV [56], which can be understood as an automatic parameter tuning. It consists of two main parts: grid search for hyperparameters and model cross-validation.

The three common tuning methods are random search, grid search and Bayesian optimization, among which grid search is the simplest and most widely used hyperparameter search method [57], [58]. The grid search is the simplest and most widely used hyperparameter search method, which determines the optimal value by searching all points in the hyperparameter range and ensuresthat the hyperparameter that makes the model most accurate within the given parameter range is found. Cross-validation [59] setsaside a part of the data in the training set as the validation set to evaluate

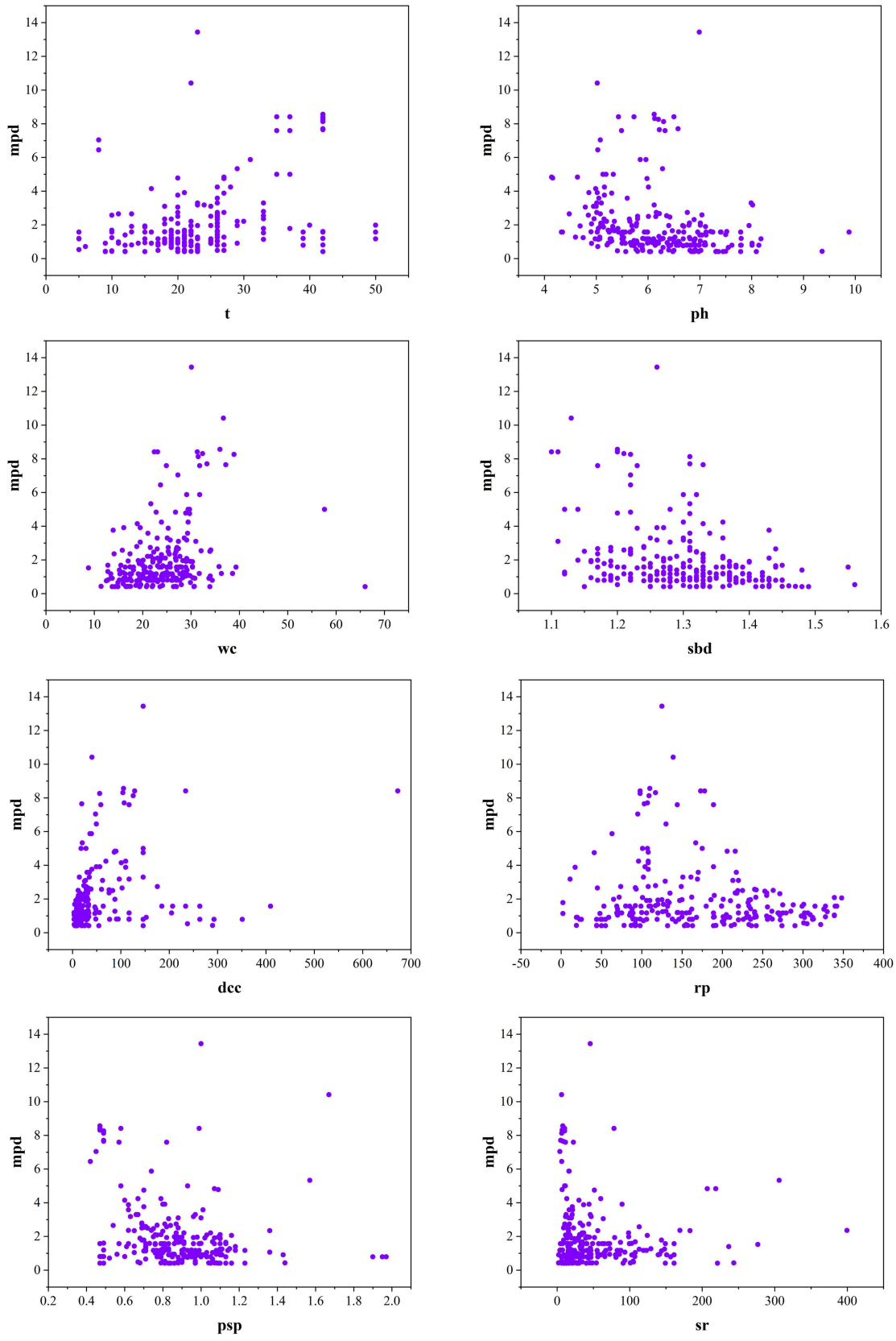

FIGURE 3. Scatter plot of the variation of each parameter with mpd.

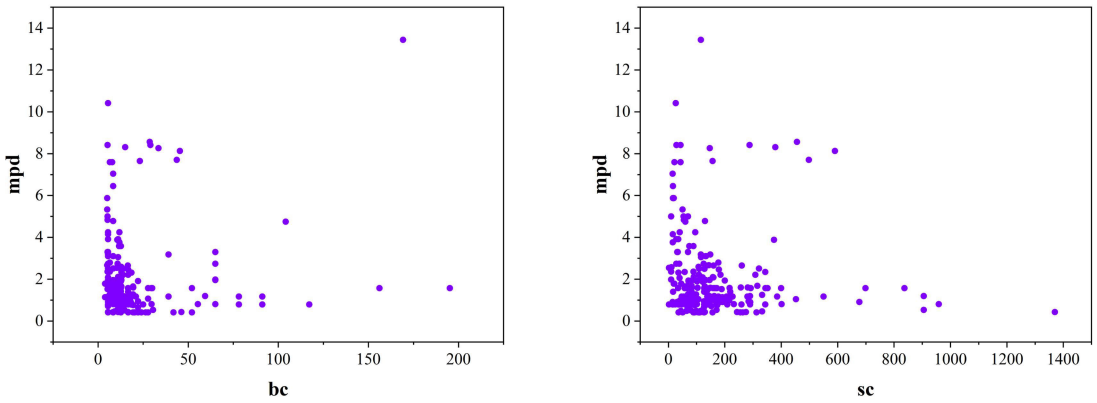

FIGURE 3. *(Continued.)* **Scatter plot of the variation of each parameter with mpd.**

TABLE 4. **The hyperparameters of the XGBoost model and its search scope.**

Hyparameters	Search scope
max_depth	[2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]
n_estimators	[20,50,80,100,110,120,150,180,200,300,500,800,1000,1500,1800.2000]
learning_rate	np.linspace(0.01, 0.3 ,30),np.linspace(0.4, 1.0, 13)
subsamples	[0.5, 0.7, 0.8, 0.9, 1]
gamma	[0, 0.1, 0.5, 1, 2, 3, 4, 5]
min_child_weight	[1, 2, 3, 4, 5]
colsample_bytree	[0.5, 0.7, 0.8, 0.9, 1]
Note: np.linspace(a, b, c) represents the number of c produced uniformly from a to b.	

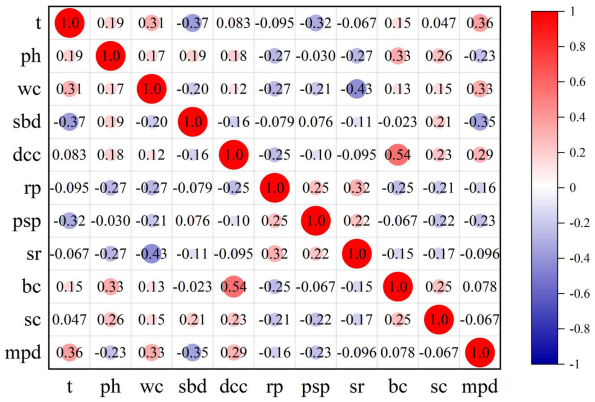

FIGURE 4. **Correlation analysis chart.**

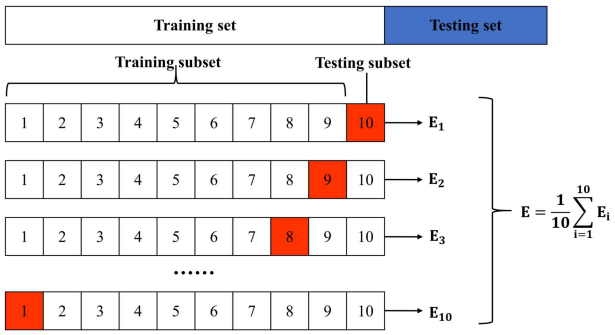

FIGURE 5. **10-fold CV.**

the model performance, selectsthe best hyperparameters if the amount of data is sufficient, and the test set makesthe final evaluation index on the screened hyperparameters. In this study, K-fold cross-validation [60], [61] is used, whichcan minimize the model bias due to random sampling training and reduce the adverse effects of data division on the model. The original data set is divided into a training set and a test set, and the training set is thendivided into k parts on average. Each training k-1 copy of the data isused as the training set and the remaining 1 copy is used as the validation set, and k models are built by training k times in total. Within a certain range, the larger the k-value, the more times the training, the more the average value taken out at the end can represent the accuracy of the model. However, the k-value cannot be increased indefinitely and must be limited to a certain range. Generally speaking, a k value of 10 is considered reasonable. In this study, we use 10 times cross-validation, that is,the k value is equal to 10.

GirdSearchCV is a training and comparison processthat combines the advantages of grid search and cross-validation to achieve automatic parameter tuning. In other words, the parameters are sequentially tuned,the learner model is trained in a given range of hyperparameters in steps, and the hyperparameter that gives the highest accuracy to the validation set can be selected from all parameters. In this study, 10-fold cross-validation is used to evaluate the performance of each combination of parameters of the XGBoost model and to

TABLE 5. The hyperparameter values of machine learning models.

Machine learning models	Hyparameters
BP	/
SVR	Kernel='rbf',C=1.0,epsilon=0.1
DTR	max_depth=3,min_samples_split=2
RF	max_depth=6,n_estimators=100
AdaBoost	n_estimators=100,learning rate=1.0
GSCV-XGBoost	n_estimators=600,max depth=7,subsamples=1, gamma=1,min_child_weight=0.9,colsample_bytree=1

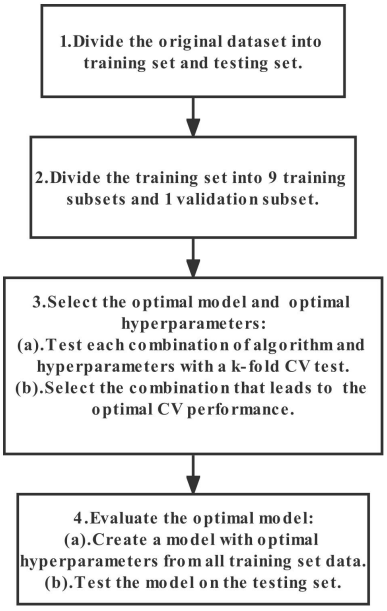

FIGURE 6. The four-step process for optimizing the XGBoost model using the GirdSearchCV method.

select the optimal combination of parameters that minimizes the model error. The selected optimal model and optimal hyperparameter combinations are used to train the entire training set and finally used on the test set for the final prediction.

Figure 5 showsthe10-fold cross-validation method. Figure 6 shows the four-step process of optimizing the XGBoost model using the GirdSearchCV method. The first step is to randomly scramble the 255 sets of pipeline datasets in this study and divide them into training and test sets, with aratio of8:2.The training set has 204 sets of data and the test set has 51 sets. The second step divides the training set into 10 equal points on average, nineof which are used as the training set and the remaining 1 part is the verification set, and 10 partitions are carried out in turn, and a total of 10 machine learning models are established. The third step is to select the combination of the optimal model hyperparameters, adjust the hyperparameters in turn according to the step size, use them to train these 10 models, and select the combination of the model and the hyperparameters that make the verification set the most accurate. The fourth step is to train the entire training set using the selected optimal model and the optimal hyperparameter combination, and finally to make the final prediction on the test set. The optimized GSCV-XGBoost model was finally obtainedthrough the GirdSearchCV optimization process.

The hyperparameter search for the XGBoost model was performed. The searched hyperparameters and their search ranges are shown in Table 4 below.

V. RESULTS AND DISCUSSION

A. MODEL PREDICTION RESULTS

Forthe BP, SVR, DTR, RF and AdaBoost models, the dataset was divided into training and test setsaccording to 8:2. The training set was used entirely for model training and the test set was used for prediction of results. The optimized GSCV-XGBoost model and the remaining five benchmark models were used for model prediction in the test set, respectively. The parameter settings for each model are listed in Table 5. Figure7 shows the final result performance of the GSCV-XGBoost model and the remaining five comparison models on the 51 sets of test sets, showing the comparison between the true value of the maximum corrosion depth mpd and the predicted output value. The red dots in the figure represent the real values of the mpd samples on the 51 test sets, the blue dots represent the predicted output values of mpd in the test sets of the model, and the solid blue line represents the line of predicted values. The fit of each model is shown in the figure. The GSCV-XGBoost model has the best prediction effect, most of the blue prediction points are close to the red true value points, and the overall difference between the true value and the predicted value is very small. The three models DTR,RF and AdaBoost fit similarly and are weaker than the GSCV-XGBoost modelis. More than half of the blue predicted points are closer to the red true value points, and the rest of the predicted points deviate more from the true values.The BP model fits weaker than these three models, with more predicted points deviating from the true points. Only a few of the predicted values in the SVR model are close to the true values, and most of the predicted value points deviate far from the red true value points. This indicates that the SVR model is the worst fit of all the models.

Figure8 shows the performance of the GSCV-XGBoost model and the remaining five benchmark models on the 51 test sets.

Each purple point in the figure represents a test set sample, and its x-coordinate corresponds to its true value and

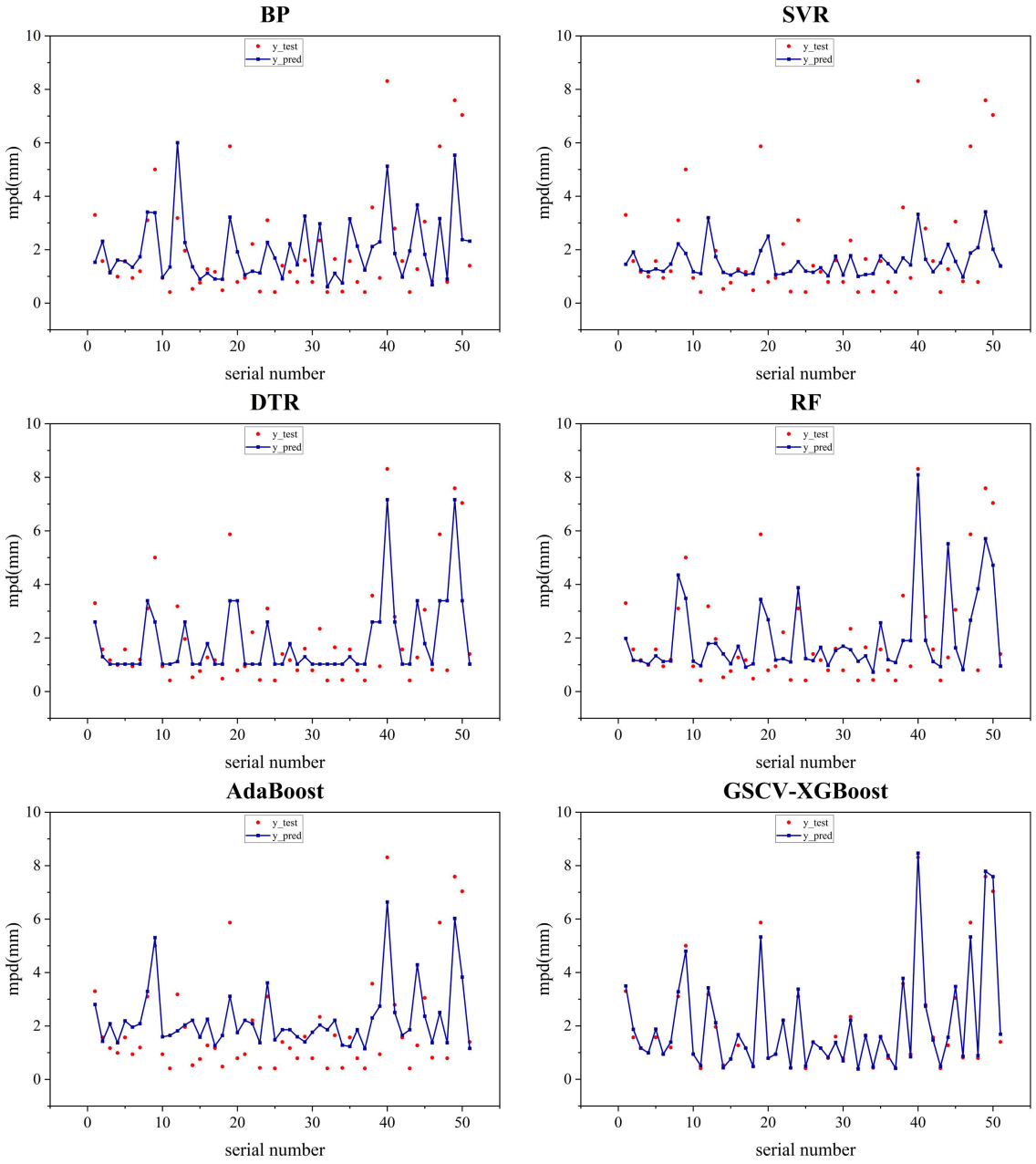

FIGURE 7. Comparison between actual and predicted values of 6 machine learning models.

y-coordinate corresponds to the model prediction value. The closer the purple dot is to the red y = x line, the smaller the difference between the true and predicted values of the sample, and the better the model fit. The dashedblackline in the figure represents the fitted curve of these sample points. The GSCV-XGBoost model has the highest degree of overlap between the sample points and the y = x line. The sample points are basically located on the line or distributed on both sides of the line, and the true value of the test set fits the predicted value to the highest degree. The DTR, AdaBoost, RF, and BP models are the next best fit, with some sample points distributed near the y = x line. Most points in the SVR model are far from the y = x line, and the fitted curves of the sample points deviate a lot from the y = x line. Moreover, the maximum corrosion depth samples larger than 4mm in the SVR model, their predicted values are all much smaller than their true values. indicating that the SVR model hasthe worst fit.

B. ERROR ASSESSMENT OF PREDICTION RESULTS

In order to accurately assess the prediction accuracy of each model, an error analysis of the final prediction of each model was conducted in this study. The calculated error metric scores for each model are listed in Table 6. R^2 and EV

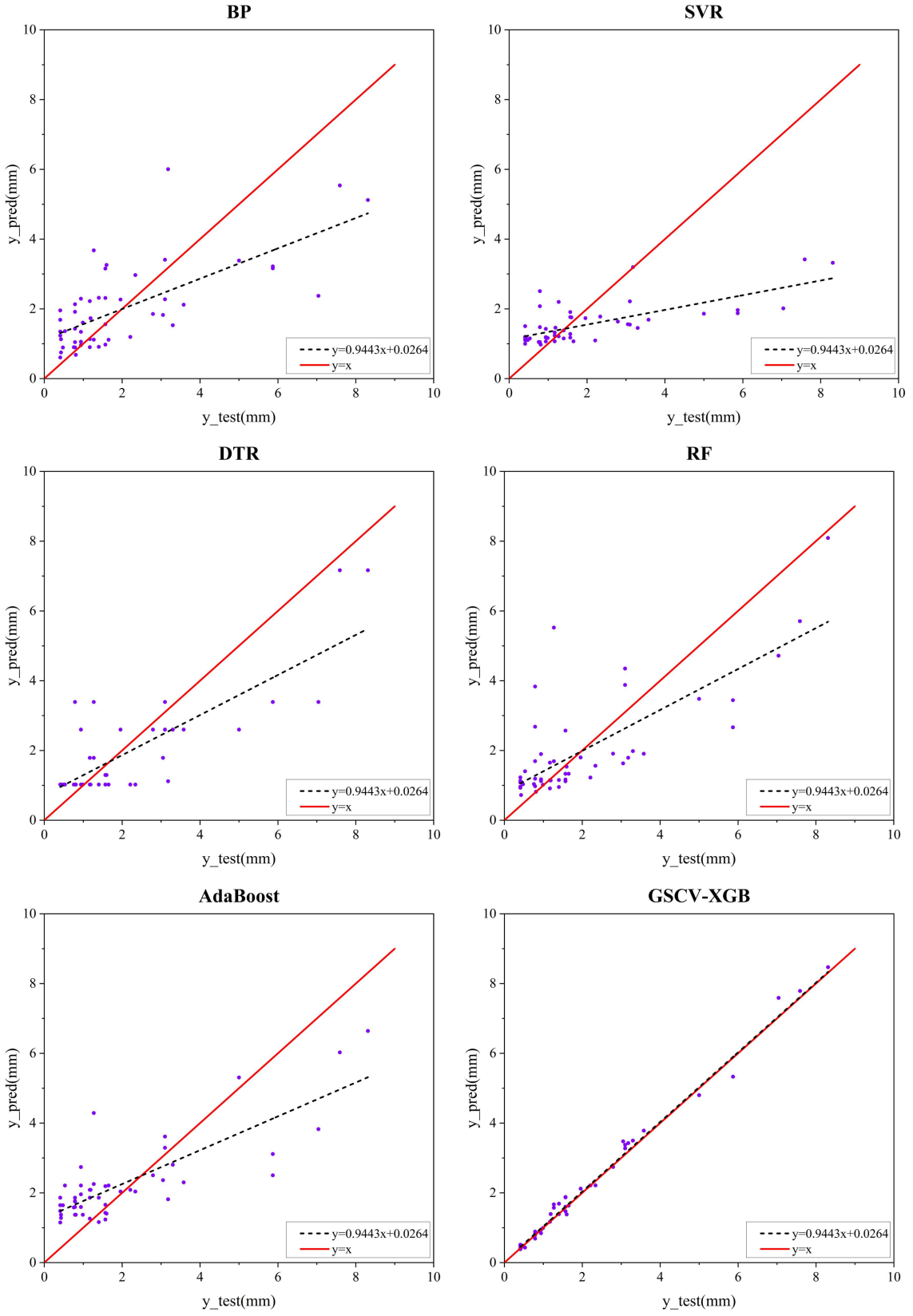

FIGURE 8. Fit of 6 machine learning models on the test set.

TABLE 6. Error metrics for 6 ML models.

	BP	SVR	DTR	RF	AdaBoost	**GSCV-XGBoost**	Velazquez et al.	Ben Seghier et al.
EV	0.4758	0.3315	0.6310	0.5778	0.6075	**0.9894**	/	/
R^2	0.4758	0.2764	0.6261	0.5778	0.5903	**0.9886**	/	/
MSE	1.9394	2.6769	1.3830	1.5618	1.5156	**0.0423**	1.1311	0.5588
RMSE	1.3926	1.6361	1.1760	1.2497	1.2311	**0.2057**	2.3153	0.2909
MAE	1.0287	1.0249	0.8304	0.8793	0.9589	**0.1401**	1.2165	0.2359
MedAE	0.8263	0.5898	0.5478	0.5509	0.8223	**0.0999**	/	/

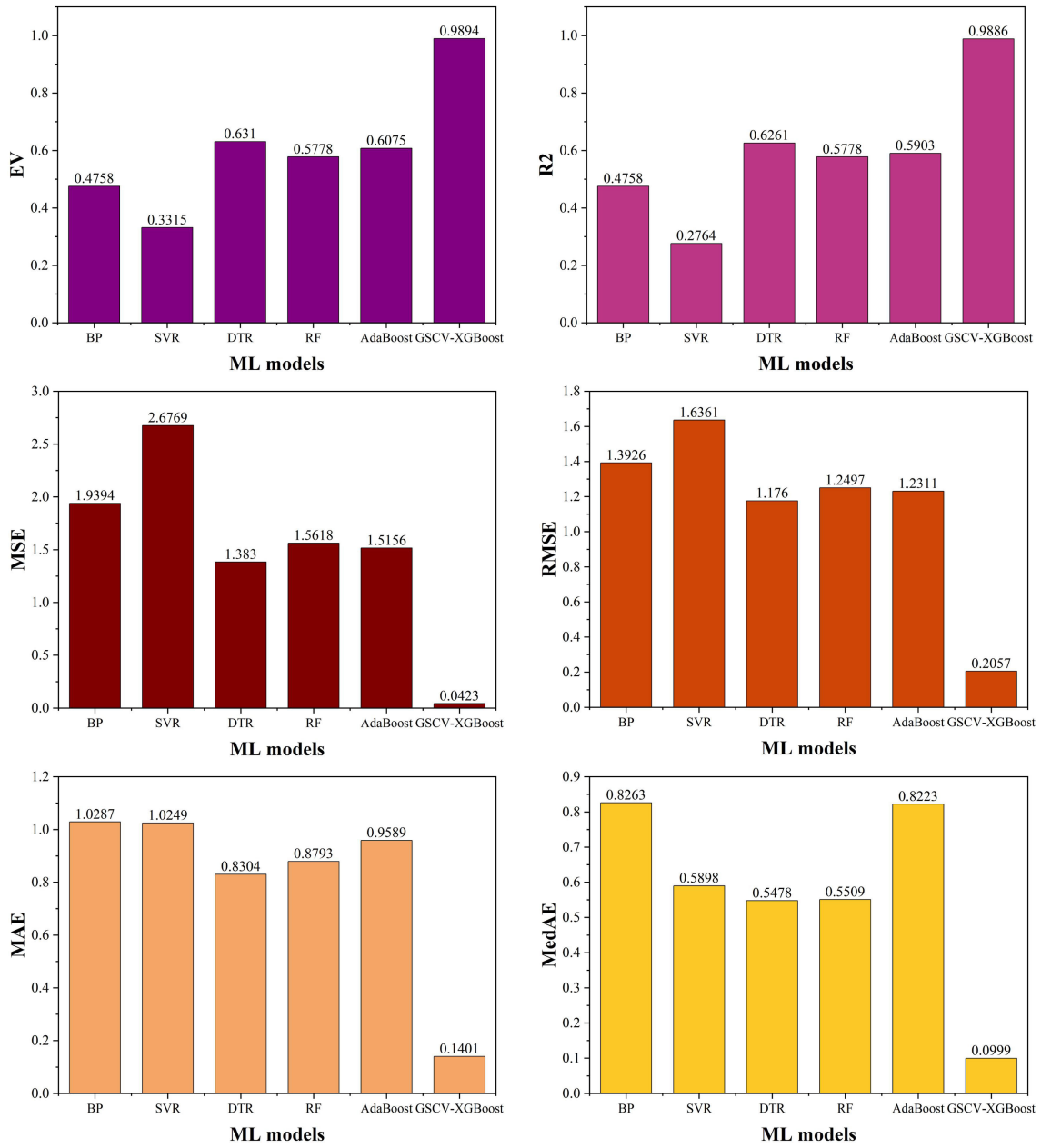

FIGURE 9. Histogram of error metric scores for 6 ML models.

represent the fitting accuraciesof the models. Among all the samples in the test set, GSCV-XGBoost has the highest accuracy with an explainable variance value EV and a coefficient of determination R^2 of 0.9894 and 0.9886, respectively. The EV and R^2 scores of DTR are 0.6310 and 0.6261, respectively. AdaBoost, RF, and BP have the next highest

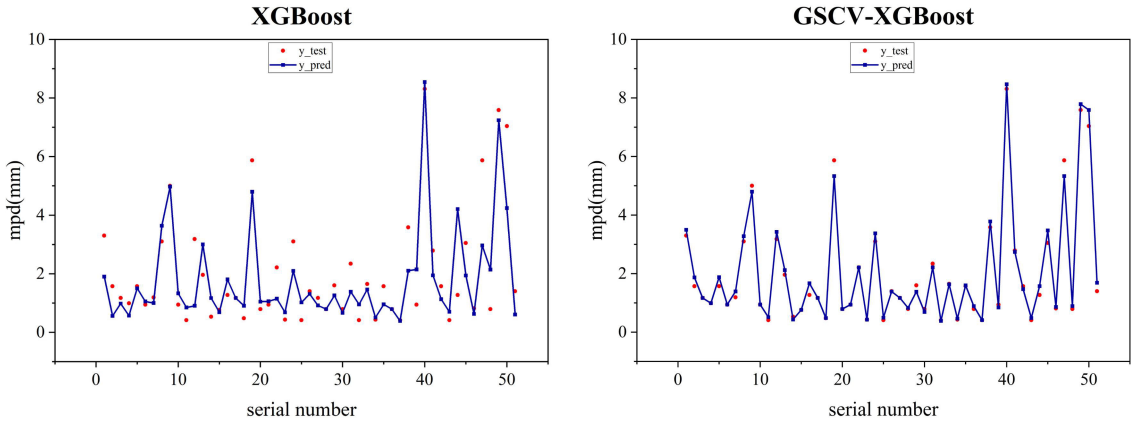

FIGURE 10. Prediction results of the XGBoost model and the GSCV-XGBoost model.

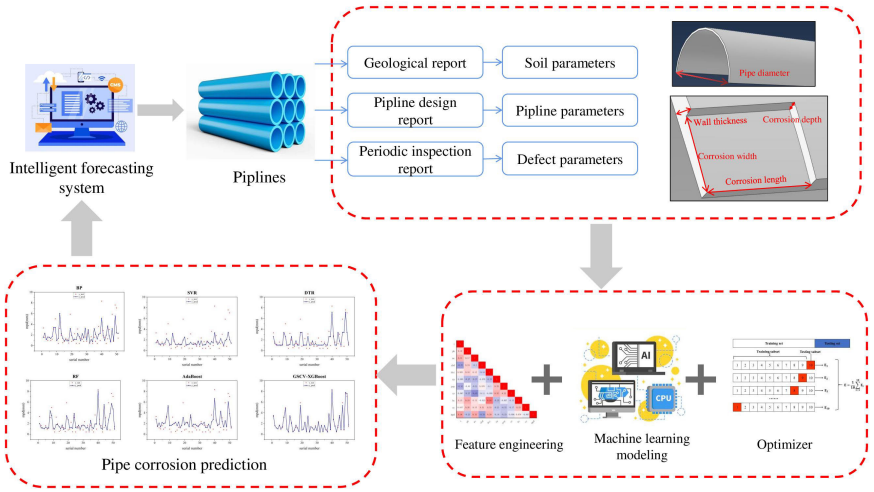

FIGURE 11. Flow chart of real-world application.

accuracy with their R^2 scores of 0.5903, 0.5778, and 0.4758, respectively. SVR has the lowest accuracy metrics.The EV and R^2 scores of the SVR model are only 0.3315 and 0.2764, respectively.

MSE, RMSE, MAE, and MedAE represent the errors of the models.The MSE, RMSE, MAE, and MedAE scores of the GSCV-XGBoost model are 0.0423, 0.2057, 0.1401, and 0.0999, respectively, and these prediction errors are the smallest among the models. The prediction errors for each of the DTR, RF, and AdaBoost models are larger than the GSCV-XGBoost model, but smaller than the BP and SVR models.The MSE,RMSE,MAE,MedAE scores of BP and SVR models are higher than other models.

It shows that GSCV-XGBoost performs the best in predicting the maximum corrosion depth dataset, and it has the highest prediction performance with the lowest error.The SVR model and the BP model have the highest prediction errorsand the SVR model has the lowest prediction performance. Figure 9 shows the histogram of the scores of each error metric parameter for the GSCV-XGBoost model and the other five compared models.

Moreover, the results of the study were compared with thoseof Ben Seghier and Velazquez et al. As shown in Table 6, the prediction errors of the GSCV-XGBoostmodel proposed in this study are smaller than those of these researchers.

C. INFLUENCE OF PARAMETER OPTIMIZATION

This study differs from other prediction models in that it uses GirdSearchCV, a parametric optimization approach, and whether parameter optimization is performed has a significant impact on the prediction accuracy of the model. To demonstrate this difference more clearly, this study specifically compares the prediction results of the XGBoost base model without parameter optimization and the GSCV-XGBoost model with parameter optimization on the test set, as shown in Figure10. The left panel shows the prediction results of the XGBoost model without parameter optimization on the test set, and the right panel shows the prediction

TABLE 7. Error indicators for the XGBoost model and the GSCV-XGBoost model.

	EV	R^2	MSE	RMSE	MAE	MedAE
XGBoost	0.7456	0.7345	0.9822	0.9911	0.6723	0.4268
GSCV-XGBoost	0.9894	0.9886	0.0423	0.2057	0.1401	0.0999

results of the GSCV-XGBoost model. The red dots indicate the real values of mpd, the blue dots indicate the predicted values of mpd, and the blue curve represents the concatenation of the predicted values. It can be clearly seen that the XGBoost model has a poor fit between the predicted and true values of the maximum corrosion depth, whereas the GSCV-XGBoost model has a significantly better fit. Table 7shows the error index scores of the XGBoost and GSCV-XGBoost models forthe test set.From Table 7, it is evident that the XGBoost model is poorly fitted, and the accuracy metrics EV and R^2 are 0.7456 and 0.7345, respectively. The various errors of the XGBoost model are also relatively large, with MSE, RMSE, MAE, and MedAE being 0.9822, 0.9911, 0.6723, and 0.4268, respectively. The accuracy of the GSCV-XGBoost model after parameter optimization is greatly improved, and the coefficient of determination R^2 is increased by 34.59%, and the root mean square error RMSE is reduced to 0.2057. This also shows that parameter optimization is crucial for the model, which can significantly improve the model prediction accuracy and reduce the model prediction error.

VI. REAL-WORLD APPLICATION

The pipeline maximum corrosion depth intelligent prediction method proposed in this studyis applied in a practical scenario with the flow shown in Figure11. Pipeline data collection should be performed first, which is an important step. The data can be obtained from geological, pipeline design, and periodic inspection reports, which constitute the data set for pipeline analysis. The acquired dataset is processed for feature engineering to extract the important parameters. A maximum corrosion depth prediction model is constructed using machine learning algorithms and the model is optimized. The residual value and probability of failure of the defective pipeline can be better assessedbased on the maximum corrosion depth of the pipeline predicted by the optimized model. Subsequently, atime prediction model of the maximum corrosion depth of the pipe can be added to better predict the growth of corrosion depth. Combined with pipeline base information, a reasonable prediction of the remaining life of the pipelinecan be made. Relevant personnel can make a reasonable pipeline cycle maintenance management plan based on the predicted value of the model. This intelligent prediction method proposed in this study can play a role in practical engineering scenarios,such as pipeline operation and maintenance management, engineering design evaluation, asset management and optimization, and safety risk assessment. Compared with traditional formula calculation and finite element simulation methods, the use of data-driven artificial intelligence algorithms can simplify the workflow and improve efficiency. Artificial intelligence algorithms can achieve even better performance in future pipeline research.

VII. CONCLUSION

In this study, the influence of various factors on the maximum corrosion depth of a pipeline is analyzed and considered, and the maximum corrosion depth prediction model of pipeline based on GSCV-XGBoost is established. A prediction framework that can be applied to practical engineering is proposed. The pipeline features were extracted by correlation analysis, and the feature variables important to the maximum corrosion depth were obtained. The XGBoost model was optimized using the GirdSearchCV method to obtain the hybrid GSCV-XGBoost model. The prediction results werecompared with those of the remaining five common machine learning models, and with the work of other researchers. The effects of parameter optimization on the prediction accuracy of the model were compared. The following conclusions were drawn.

The GSCV-XGBoost model has the highest accuracy in predicting the maximum corrosion depth of a pipeline. This study testedthe prediction accuracy of the improved GSCV-XGBoost algorithm compared with the remaining five common machine learning algorithms on the pipeline dataset. First, the control variables of the maximum corrosion depth of the pipeline were processed using Pearson correlation analysis, and the input variables of the model wereobtained through feature extraction. Second, the basic machine learning model XGBoost was improved. The GirdSearchCV algorithm wasusedto tune the hyperparameters of the traditional machine learning model. A 10-fold CV wasused to reduce the adverse effects caused by random division of data during model training. Simultaneously, the optimal hyperparameter combination of the model was obtained by combining the grid search to obtain the optimized GSCV-XGBoost model. Finally, the GSCV-XGBoost model and the remaining five comparison models wereused to make predictions forthe pipeline test set, and the prediction results wereevaluated for errors. The results show that the prediction model based on GSCV-XGBoost has the lowest error and highest prediction accuracy among all the models. The prediction results are significantly better than those of other researchers. The RMSEof the GSCV-XGBoost model is only 0.2057, and the EV and R^2, which represent the prediction accuracy, are 0.9894and 0.9886, respectively. It has the absolute advantage of high accuracy compared with other models.The GSCV-XGBoost model has the highest fit and strongest predictive ability for the maximum pipeline corrosion depth dataset. All the other models suffered from underfitting or overfitting problems. Through the correlation analysis of the data set, we also found that the parameters that have the greatest influence on the maximum corrosion depth are the exposure time of the pipeline, soil accumulation density, and soil moisture content.

Parameter optimization has an important impact on the prediction accuracy of machine learning models. In this study, we comparedand analyzedthe impact of with and

without parameter optimization on the prediction results of the XGBoost models. The comparison results show that the baseXGBoost model without GirdSearchCV optimized parameters has a high RMSE of 0.9911, and an R^2 score of only 0.7345 forthe test set. The hybrid model GSCV-XGBoost after parameter optimization is much better than the basic XGBoost model in terms of prediction accuracy. The accuracy metric R^2 improved by 34.59% and the RMSE decreased to 0.2057. This indicates that the prediction performance of the model can be better explored by GirdSearchCV optimization, which largely improves the accuracy of the model.

This GSCV-XGBoost-based intelligent prediction model proposed in this studycan refine the machine learning dataset using feature engineering and optimize the model parameters by GirdSearchCV. It has a high prediction accuracy forthe problem of maximum corrosion depth prediction of the pipelines. However, there are still shortcomings in this model, the data of this model come from the existing actual cases, and the data samples are mostly concentrated in the soil samples around the service pipeline. Many other variables affect the maximum corrosion depth of the pipeline, such as the type of steel, protection methods, and corrosives.

The specific effects of these factors on the depth of corrosion of the pipeline areasfollows. Type of steel: Different types of steelshave different corrosion resistances. Generally, steelscontaining chromium, nickel, molybdenum and other alloy elements exhibit better corrosion resistance. Protection methods: Oil and gas pipelines are typically protected against corrosion by cathodic protection. Cathodic protection technology typically involves the installationofone or more cathodes on the pipelinesurface. Through the action of the current to reduce the pipeline surface potential to below the cathodic potential, the metal surface becomes a cathode to achieve anti-corrosion. Corrosive agents: There are many corrosive agents in the working environment of oil and gas pipelines, such as oxygen, water vapor, carbon dioxideand hydrogen sulfide. The concentration of these corrosives and their time of action also affect the corrosiondepth. For example, a high concentration of hydrogen sulfide will accelerate the corrosion of steel, whereas carbon dioxide slowsdown the corrosion rate to a certain extent. The inclusion of more influencing factors should be considered in future studiesbecause of the lack of these influencing factors. Deep learning was not applied in this study, and a better model can be built in the future by applying deep learning algorithms. In addition, the corrosion depth growth trend of the pipeline and the probability of failure of the pipeline can be investigated in futurestudies.

REFERENCES

[1] M. Guo, X. Zhu, G. Huang, M. Li, and Z. Li, "Analysis on standardization and informatization development of oil and gas pipeline full life cycle management," *Mod. Chem. Ind.*, vol. 42, no. 5, pp. 14–18, 2022.

[2] L. Xu, Y. Wang, L. Mo, Y. Tang, F. Wang, and C. Li, "The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines," *Eng. Failure Anal.*, vol. 144, Feb. 2023, Art. no. 106951, doi: 10.1016/j.engfailanal.2022.106951.

[3] H. Zhang, Q. Feng, B. Yan, X. Zheng, Y. Yang, J. Chen, H. Zhang, and X. Liu, "State of the art of oil and gas pipeline vulnerability assessments," *Energies*, vol. 16, no. 8, p. 3439, Apr. 2023, doi: 10.3390/en16083439.

[4] W. Wang, J. Hu, X. Yuan, L. Zhou, J. Yu, Z. Zhang, and X. Zhong, "Understanding the effect of tensile stress on erosion-corrosion of X70 pipeline steel," *Construct. Building Mater.*, vol. 342, Aug. 2022, Art. no. 127972, doi: 10.1016/j.conbuildmat.2022.127972.

[5] J. Fang, X. Cheng, H. Gai, S. Lin, and H. Lou, "Development of machine learning algorithms for predicting internal corrosion of crude oil and natural gas pipelines," *Comput. Chem. Eng.*, vol. 177, Sep. 2023, Art. no. 108358, doi: 10.1016/j.compchemeng.2023.108358.

[6] J. C. Velázquez, E. Hernández-Sánchez, G. Terán, S. Capula-Colindres, M. Diaz-Cruz, and A. Cervantes-Tobón, "Probabilistic and statistical techniques to study the impact of localized corrosion defects in oil and gas pipelines: A review," *Metals*, vol. 12, no. 4, p. 576, Mar. 2022, doi: 10.3390/met12040576.

[7] G. Qin, Y. Huang, Y. Wang, and Y. Frank Cheng, "Pipeline condition assessment and finite element modeling of mechano-electrochemical interaction between corrosion defects with varied orientations on pipelines," *Tunnelling Underground Space Technol.*, vol. 136, Jun. 2023, Art. no. 105101, doi: 10.1016/j.tust.2023.105101.

[8] X. Miao, H. Zhao, B. Gao, and F. Song, "Corrosion leakage risk diagnosis of oil and gas pipelines based on semi-supervised domain generalization model," *Rel. Eng. Syst. Saf.*, vol. 238, Oct. 2023, Art. no. 109486, doi: 10.1016/j.ress.2023.109486.

[9] Y. Wang, R. Li, A. Xia, P. Ni, and G. Qin, "An integrated modeling method of uncertainties: Application-orientated fuzzy random spatiotemporal analysis of pipeline structures," *Tunnelling Underground Space Technol.*, vol. 131, Jan. 2023, Art. no. 104825, doi: 10.1016/j.tust.2022.104825.

[10] F. Caleyo, J. C. Velázquez, A. Valor, and J. M. Hallen, "Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study," *Corrosion Sci.*, vol. 51, no. 9, pp. 1925–1934, Sep. 2009, doi: 10.1016/j.corsci.2009.05.019.

[11] Z. Luo, Y. Song, and A. Bi, "Prediction model of internal corrosion rate in oil and gas gathering pipelines based on GRA-RFR," *Mater. Protection*, vol. 53, no. 3, pp. 95–100, 2020.

[12] D. P. Yang, Y. M. Wang, Y. F. Cao, F. L. Long, and G. Q. Niu, "Study on residual strength of corroded pipes," *Mech. Eng., Mater. Inf. Technol. II*, vol. 662, p. 196, 2014, doi: 10.4028/www.scientific.net/AMM.662.196.

[13] J. Yang, Y. Li, and Y. Peng, "Numerical analysis on welding residual stress in Q690 high-strength steel pipe," *Eng. Mech.*, vol. 31, no. 10, pp. 108–115, 2014.

[14] C.-L. Su, X. Li, and J. Zhou, "Failure pressure analysis of corroded moderate-to-high strength pipelines," *China Ocean Eng.*, vol. 30, no. 1, pp. 69–82, Mar. 2016, doi: 10.1007/s13344-016-0004-z.

[15] H. Seo, J. No, and S. S. Park, "ml-SFP: System failure prediction method based on machine learning," in *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022*, vol. 2. 2023, pp. 195–203.

[16] D. Ma, H. Fang, N. Wang, C. Zhang, J. Dong, and H. Hu, "Automatic detection and counting system for pavement cracks based on PCGAN and YOLO-MF," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22166–22178, Nov. 2022, doi: 10.1109/TITS.2022.3161960.

[17] S. D. V. Kumar, S. Karuppanan, and M. Ovinis, "Artificial neural network-based failure pressure prediction of API 5L X80 pipeline with circumferentially aligned interacting corrosion defects subjected to combined loadings," *Materials*, vol. 15, no. 6, p. 2259, Mar. 2022, doi: 10.3390/ma15062259.

[18] M. Lo, S. Karuppanan, and M. Ovinis, "Failure pressure prediction of a corroded pipeline with longitudinally interacting corrosion defects subjected to combined loadings using FEM and ANN," *J. Mar. Sci. Eng.*, vol. 9, no. 3, p. 281, Mar. 2021, doi: 10.3390/jmse9030281.

[19] N. B. Shaik, S. R. Pedapati, S. A. A. Taqvi, A. R. Othman, and F. A. A. Dzubir, "A feed-forward back propagation neural network approach to predict the life condition of crude oil pipeline," *Processes*, vol. 8, no. 6, p. 661, Jun. 2020, doi: 10.3390/pr8060661.

[20] K. F. Tee and A. H. Wordu, "Burst strength analysis of pressurized steel pipelines with corrosion and gouge defects," *Eng. Failure Anal.*, vol. 108, Jan. 2020, Art. no. 104347, doi: 10.1016/j.engfailanal.2019.104347.

[21] H. Fang, K. Yang, B. Li, P. Tan, F. Wang, and X. Du, "Experimental and numerical study on mechanical analysis of buried corroded concrete pipes under static traffic loads," *Appl. Sci.*, vol. 9, no. 23, p. 5002, Nov. 2019, doi: 10.3390/app9235002.

[22] W. Wang, W. Shi, and C.-Q. Li, "Time dependent reliability analysis for cast iron pipes subjected to pitting corrosion," *Int. J. Pressure Vessels Piping*, vol. 175, Aug. 2019, Art. no. 103935, doi: 10.1016/j.ijpvp.2019.103935.

[23] B. Ma, J. Shuai, D. Liu, and K. Xu, "Assessment on failure pressure of high strength pipeline with corrosion defects," *Eng. Failure Anal.*, vol. 32, pp. 209–219, Sep. 2013, doi: 10.1016/j.engfailanal.2013.03.015.
[24] Y. Chen, D. Zhang, Y. Wang, and S. Xu, "Corrosion pit depth prediction model of nuclear power pipeline using fractal theory," *At. Energy Sci. Technol.*, vol. 43, no. 8, pp. 673–677, 2009.
[25] H. F. Lu, H. Y. Peng, Z. D. Xu, J. C. Matthews, N. N. Wang, and T. Iseley, "A feature selection-based intelligent framework for predicting maximum depth of corroded pipeline defects," *J. Perform. Constructed Facilities*, vol. 36, no. 5, 2022, Art. no. 04022044, doi: 10.1061/(ASCE)CF.1943-5509.0001753.
[26] G. Ma, J. Li, R. Bai, and Z. Dai, "Prediction of corrosion rate in oil and gas pipelines based on PSO-SVM model," *Surf. Technol.*, vol. 48, no. 5, pp. 43–48, 2019.
[27] N. Balekelayi and S. Tesfamariam, "External corrosion pitting depth prediction using Bayesian spectral analysis on bare oil and gas pipelines," *Int. J. Pressure Vessels Piping*, vol. 188, Dec. 2020, Art. no. 104224, doi: 10.1016/j.ijpvp.2020.104224.
[28] M. El Amine Ben Seghier, B. Keshtegar, K. F. Tee, T. Zayed, R. Abbassi, and N. T. Trung, "Prediction of maximum pitting corrosion depth in oil and gas pipelines," *Eng. Failure Anal.*, vol. 112, May 2020, Art. no. 104505, doi: 10.1016/j.engfailanal.2020.104505.
[29] J. C. Velázquez, F. Caleyo, A. Valor, and J. M. Hallen, "Technical note: Field study—Pitting corrosion of underground pipelines related to local soil and pipe characteristics," *Corrosion*, vol. 66, no. 1, pp. 016001-1–016001-5, Jan. 2010, doi: 10.5006/1.3318290.
[30] J. C. Velázquez, F. Caleyo, A. Valor, and J. M. Hallen, "Predictive model for pitting corrosion in buried oil and gas pipelines," *Corrosion*, vol. 65, no. 5, pp. 332–342, May 2009, doi: 10.5006/1.3319138.
[31] C. I. Ossai, "A data-driven machine learning approach for corrosion risk assessment—A comparative study," *Big Data Cognit. Comput.*, vol. 3, no. 2, p. 28, May 2019, doi: 10.3390/bdcc3020028.
[32] R. K. Mazumder, A. M. Salman, and Y. Li, "Failure risk analysis of pipelines using data-driven machine learning algorithms," *Struct. Saf.*, vol. 89, Mar. 2021, Art. no. 102047, doi: 10.1016/j.strusafe.2020.102047.
[33] Y. Zhu, J. Jiang, L. Zhao, H. Liu, M. Hou, and H. Li, "Survey on application of correlation analysis of different calculation forms in meteorology," *J. Trop. Meteorol.*, vol. 37, no. 1, pp. 1–13, 2021.
[34] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 3, p. e1301, May 2019, doi: 10.1002/widm.1301.
[35] A. Akl, I. El-Henawy, A. Salah, and K. Li, "Optimizing deep neural networks hyperparameter positions and values," *J. Intell. Fuzzy Syst.*, vol. 37, no. 5, pp. 6665–6681, Nov. 2019, doi: 10.3233/JIFS-190033.
[36] M. A. Salam, L. Ibrahim, and D. S. Abdelminaam, "Earthquake prediction using hybrid machine learning techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, pp. 654–665, 2021.
[37] A. A. Shah, K. Ahmed, X. Han, and A. Saleem, "A novel prediction error-based power forecasting scheme for real PV system using PVUSA model: A grey box-based neural network approach," *IEEE Access*, vol. 9, pp. 87196–87206, 2021, doi: 10.1109/ACCESS.2021.3088906.
[38] M. Parsa, "A data augmentation approach to XGboost-based mineral potential mapping: An example of carbonate-hosted Zn-Pb mineral systems of western Iran," *J. Geochemical Explor.*, vol. 228, Sep. 2021, Art. no. 106811, doi: 10.1016/j.gexplo.2021.106811.
[39] X. Zhang, H. Nguyen, X.-N. Bui, Q.-H. Tran, D.-A. Nguyen, D. T. Bui, and H. Moayedi, "Novel soft computing model for predicting blast-induced ground vibration in open-pit mines based on particle swarm optimization and XGBoost," *Natural Resour. Res.*, vol. 29, no. 2, pp. 711–721, Apr. 2020, doi: 10.1007/s11053-019-09492-7.
[40] H. Nguyen, X.-N. Bui, H.-B. Bui, and D. T. Cuong, "Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: A case study," *Acta Geophys.*, vol. 67, no. 2, pp. 477–490, Apr. 2019, doi: 10.1007/s11600-019-00268-4.
[41] W. Alajali, W. Zhou, S. Wen, and Y. Wang, "Intersection traffic prediction using decision tree models," *Symmetry*, vol. 10, no. 9, p. 386, Sep. 2018, doi: 10.3390/sym10090386.
[42] J. Su, Y. Wang, X. Niu, S. Sha, and J. Yu, "Prediction of ground surface settlement by shield tunneling using XGBoost and Bayesian optimization," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105020, doi: 10.1016/j.engappai.2022.105020.
[43] L. Chen, F. Zhang, and L. Sun, "Research on the calibration of binocular camera based on BP neural network optimized by improved genetic simulated annealing algorithm," *IEEE Access*, vol. 8, pp. 103815–103832, 2020, doi: 10.1109/ACCESS.2020.2992652.
[44] G. Du, Z. Liu, and H. Lu, "Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment," *J. Comput. Appl. Math.*, vol. 386, Apr. 2021, Art. no. 113260, doi: 10.1016/j.cam.2020.113260.
[45] N. Manju, C. M. Samiha, S. P. P. Kumar, H. L. Gururaj, and F. Flammini, "Prediction of aptamer protein interaction using random forest algorithm," *IEEE Access*, vol. 10, pp. 49677–49687, 2022, doi: 10.1109/ACCESS.2022.3172278.
[46] A. R. Linero, "Bayesian regression trees for high-dimensional prediction and variable selection," *J. Amer. Stat. Assoc.*, vol. 113, no. 522, pp. 626–636, Apr. 2018, doi: 10.1080/01621459.2016.1264957.
[47] H. S. Barjouei, H. Ghorbani, N. Mohamadian, D. A. Wood, S. Davoodi, J. Moghadasi, and H. Saberi, "Prediction performance advantages of deep machine learning algorithms for two-phase flow rates through wellhead chokes," *J. Petroleum Explor. Prod. Technol.*, vol. 11, no. 3, pp. 1233–1261, Mar. 2021, doi: 10.1007/s13202-021-01087-4.
[48] H. Ghorbani, D. A. Wood, A. Choubineh, A. Tatar, P. G. Abarghoyi, M. Madani, and N. Mohamadian, "Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared," *Petroleum*, vol. 6, no. 4, pp. 404–414, Dec. 2020.
[49] X. Wang, X. Gao, Y. Zhang, X. Fei, Z. Chen, J. Wang, Y. Zhang, X. Lu, and H. Zhao, "Land-cover classification of coastal wetlands using the RF algorithm for worldview-2 and landsat 8 images," *Remote Sens.*, vol. 11, no. 16, p. 1927, Aug. 2019, doi: 10.3390/rs11161927.
[50] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A comparison of decision tree ensemble creation techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 173–180, Jan. 2007, doi: 10.1109/TPAMI.2007.250609.
[51] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, and H. Hong, "Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping," *Catena*, vol. 187, Apr. 2020, Art. no. 104396, doi: 10.1016/j.catena.2019.104396.
[52] Y. Cao, Q.-G. Miao, J.-C. Liu, and L. Gao, "Advance and prospects of AdaBoost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, Mar. 2014.
[53] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Nov. 2018, doi: 10.1145/3136625.
[54] L. Hussain, A. Ali, S. Rathore, S. Saeed, A. Idris, M. U. Usman, M. A. Iftikhar, and D. Y. Suh, "Applying Bayesian network approach to determine the association between morphological features extracted from prostate cancer images," *IEEE Access*, vol. 7, pp. 1586–1601, 2019, doi: 10.1109/ACCESS.2018.2886644.
[55] X. Sun, Z. Shi, G. Lei, Y. Guo, and J. Zhu, "Multi-objective design optimization of an IPMSM based on multilevel strategy," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 139–148, Jan. 2021, doi: 10.1109/TIE.2020.2965463.
[56] Y. Shuai, Y. Zheng, and H. Huang, "Hybrid software obsolescence evaluation model based on PCA-SVM-GridSearchCV," in *Proc. IEEE 9th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2018, pp. 449–453.
[57] S. W. Nurdian, N. Adu, I. R. Palupi, and W. Raharjo, "Comparison tomography relocation hypocenter grid search and guided grid search method in Java island," *J. Phys., Conf. Ser.*, vol. 776, Aug. 2016, Art. no. 012113.
[58] S. Sundhararajan, A. Pahwa, and P. Krishnaswami, "A comparative analysis of genetic algorithms and directed grid search for parametric optimization," *Eng. With Comput.*, vol. 14, no. 3, pp. 197–205, Sep. 1998, doi: 10.1007/BF01215973.
[59] J. Josse and F. Husson, "Selecting the number of components in principal component analysis using cross-validation approximations," *Comput. Statist. Data Anal.*, vol. 56, no. 6, pp. 1869–1879, Jun. 2012, doi: 10.1016/j.csda.2011.11.012.
[60] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.
[61] C. Zhang, G. Hu, D. Yurchenko, P. Lin, S. Gu, D. Song, H. Peng, and J. Wang, "Machine learning based prediction of piezoelectric energy harvesting from wake galloping," *Mech. Syst. Signal Process.*, vol. 160, Nov. 2021, Art. no. 107876, doi: 10.1016/j.ymssp.2021.107876.

NIANNIAN WANG was born in Henan, China, in 1989. She received the Ph.D. degree in structural engineering from the Dalian University of Technology, China, in 2019.

She is currently a Professor with the School of Water Resources and Civil Engineering, Zhengzhou University, China. She also serves as the Vice Chairperson of the Pipeline Inspection and Rehabilitation Committee of the China Municipal Engineering Association, and a member of the Expert Committee of the Sino-U.S. Joint Trenchless Engineering Research Center. Her research interests include the intelligent inspection of engineering structures, damage identification and assessment, and big data analysis.

Prof. Wang has led to more than ten research projects, including the National Key Research and Development Program, National Natural Science Youth Fund, China Postdoctoral Special Fund, and China Postdoctoral Fund. Her research results were awarded the Gold Prize of the First National Postdoctoral Innovation and Entrepreneurship Competition, the Second Prize of Scientific and Technological Progress in Hubei Province and Henan Province, and the ''Young Star'' of China International Trenchless Technology Symposium. She has published one monograph and 32 academic papers (25 SCI/EI papers). She was awarded the 2019 ASCE Best Paper Award (the only award-winning paper from a university in Mainland China).

LIUYANG SONG was born in Henan, China, in 1999. She received the Graduate degree from the School of Water Resources, North China University of Water Resources and Hydroelectricity, in 2017. She is currently pursuing the master's degree with the School of Water Resources and Civil Engineering, Zhengzhou University.

Her research interests include the combination of machine learning and pipeline engineering. She is primarily engaged in pipeline infrastructure disease detection and research.

HONGYUAN FANG was born in Henan, China, in 1982. He received the Ph.D. degree from the Department of Construction Engineering, Dalian University of Technology, in 2012.

He is currently a Professor and a Ph.D. Supervisor with the School of Water Resources and Civil Engineering, Zhengzhou University. He was selected as a Young Changjiang Scholar of the Ministry of Education. He is also the Vice President of the International Institute of Pipeline Professionals (IIUS), the Deputy Secretary General of the Underground Pipeline Committee of China Municipal Engineering Association, the Director of the China Trenchless Technology Association, and the Director of Henan Civil Engineering Society. His research interests include the theory and technology of nondestructive testing and the trenchless repair of water resources, transportation, and municipal infrastructure facilities.

Prof. Fang has presided over and completed more than ten national and provincial level scientific research projects in recent years, such as National Key Research and Development Program, National Natural Science Foundation of China, China Postdoctoral Science Foundation Special Grant, and Major Science and Technology Special Project of Henan Province.

BIN LI was born in Henan, China, in 1993. He received the Ph.D. degree from the Dalian University of Technology.

He is currently a member of the Chinese Society of Rock Mechanics and Engineering. His current research interests include the safe operation and maintenance of underground water supplies and drainage pipes.

FUMING WANG was born in Henan, China, in 1957. He received the Ph.D. degree from the Dalian University of Technology.

He is currently the Dean of the School of Civil Engineering, Sun Yat-sen University, the Director of the National Local Joint Engineering Laboratory of Major Infrastructure Inspection and Repair Technology, and the Director of the Henan Provincial Collaborative Innovation Center of Water Resources and Transportation Infrastructure Safety Protection. He is mainly engaged in research on infrastructure flood damage prevention and control and trenchless repair technology.

Prof. Wang received the National Technical Invention Second Class Award, the National Science and Technology Progress Second Class Award, and the National Science and Technology Progress Third Class Award, the International Trenchless Academic Research Award, and the Henan Province Science and Technology Outstanding Contribution Award. He was elected as a member of the Chinese Academy of Engineering, in 2015.