ARTICLE OPEN

Check for updates

Interpretable machine learning for maximum corrosion depth and influence factor analysis

Yuhui Song[1], Qinying Wang[1]✉, Xingshou Zhang[1], Lijin Dong[1]✉, Shulin Bai[2], Dezhi Zeng[3], Zhi Zhang[3], Huali Zhang[4] and Yuchen Xi[1]✉

We have employed interpretable methods to uncover the black-box model of the machine learning (ML) for predicting the maximum pitting depth (**dmax**) of oil and gas pipelines. Ensemble learning (EL) is found to have higher accuracy compared with several classical ML models, and the determination coefficient of the adaptive boosting (AdaBoost) model reaches 0.96 after optimizing the features and hyperparameters. In this work, the running framework of the model was clearly displayed by visualization tool, and Shapley Additive exPlanations (SHAP) values were used to visually interpret the model locally and globally to help understand the predictive logic and the contribution of features. Furthermore, the accumulated local effect (ALE) successfully explains how the features affect the corrosion depth and interact with one another.

npj Materials Degradation (2023)7:9 ; https://doi.org/10.1038/s41529-023-00324-x

INTRODUCTION

The current global energy structure is still extremely dependent on oil and natural gas resources[1]. Metallic pipelines (e.g. X80, X70, X65) are widely used around the world as the fastest, safest, and cheapest way to transport oil and gas[2–6]. Nevertheless, pipelines may face leaks, bursts, and ruptures during serving and cause environmental pollution, economic losses, and even casualties[7]. Most investigations evaluating different failure modes of oil and gas pipelines show that corrosion is one of the most common causes and has the greatest negative impact on the degradation of oil and gas pipelines[2]. Lam's[8] analysis indicated that external corrosion is the main form of corrosion failure of pipelines. External corrosion of oil and gas pipelines is a time-varying damage mechanism, the degree of which is strongly dependent on the service environment of the pipeline (soil properties, water, gas, etc.), age, and whether and how external protection is applied[1]. All of these features contribute to the evolution and growth of various types of corrosion on pipelines. Among all corrosion forms, localized corrosion (pitting) tends to be of high risk. Therefore, estimating the maximum depth of pitting corrosion accurately allows operators to analyze and manage the risks better in the transmission pipeline system and to plan maintenance accordingly.

In recent years, many scholars around the world have been actively pursuing corrosion prediction models, which involve atmospheric corrosion, marine corrosion, microbial corrosion, etc.[9–13]. To predict the corrosion development of pipelines accurately, scientists are committed to constructing corrosion models from multidisciplinary knowledge. Initially, these models relied on empirical or mathematical statistics to derive correlations, and gradually incorporated more factors and deterioration mechanisms. The increases in computing power have led to a growing interest among domain experts in high-throughput computational simulations and intelligent methods. It is a trend in corrosion prediction to explore the relationship between corrosion (corrosion rate or maximum pitting depth) and various influence factors using intelligent algorithms.

ML has been successfully applied for the corrosion prediction of oil and gas pipelines. Ren et al.[14] took the mileage, elevation difference, inclination angle, pressure, and Reynolds number of the natural gas pipelines as input parameters and the maximum average corrosion rate of pipelines as output parameters to establish a back propagation neural network (BPNN) prediction model. The predicted values and the real pipeline corrosion rate are highly consistent with an error of less than 0.1%. Xie et al.[15] and Liao et al.[16] employed the BPNN to predict the growth of corrosion in pipelines with different inputs. Meanwhile, other neural network (DNN, SSCN, et al.) models were widely used to predict corrosion of pipelines as well[17–22]. Support vector machine (SVR) is also widely used for the corrosion prediction of pipelines. Luo et al.[23] established the corrosion prediction model of the wet natural gas gathering and transportation pipeline based on the SVR, BPNN, and multiple regression, respectively. Compared with the the actual data, the average relative error of the corrosion rate obtained by SVM is 11.16%, but 19.54% and 25.32% are obtained by the ANN and multivariate analysis methods, respectively. Zhang et al.[24] combined modified SVM with unequal interval model to predict the corrosion depth of gathering gas pipelines, and the prediction relative error was only 0.82%. In addition, El Amine et al.[2] proposed an efficient hybrid intelligent model based on the feasibility of SVR to predict the **dmax** of offshore oil and gas pipelines. Although the single ML model has proven to be effective, high-performance models are constantly being developed.

EL is a composite model, and its prediction accuracy is higher than other single models[25]. Ben et al.[25] developed corrosion prediction models based on four EL approaches. In addition, they performed a rigorous statistical and graphical analysis of the predicted internal corrosion rate to evaluate the model's performance and compare its capabilities. AdaBoost and Gradient boosting (XGBoost) models showed the best performance with

[1]School of New Energy and Materials, Southwest Petroleum University, Chengdu 610500, China. [2]School of Materials Science and Engineering, Peking University, Beijing 100871, China. [3]Petroleum Engineering School, Southwest Petroleum University, Chengdu 610500, China. [4]Engineering Technology Research Institute, PetroChina Southwest Oil & Gas field Company, Guanghan 618300, China. ✉email: wangqy0401@swpu.edu.cn; ljdong89@163.com; xycsony3@126.com

RMSE values of 0.052. However, these studies fail to emphasize the interpretability of their models. Despite the high accuracy of the predictions, many ML models are uninterpretable and users are not aware of the underlying inference of the predictions[26]. That is, the prediction process of the ML model is like a black box that is difficult to understand, especially for the people who are not proficient in computer programs. Interpretable ML solves the interpretation issue of earlier models. It converts black box type models into transparent models, exposing the underlying reasoning, clarifying how ML models provide their predictions, and revealing feature importance and dependencies[27]. In recent studies, SHAP and ALE have been used for post hoc interpretation based on ML predictions in several fields of materials science[28,29]. However, how the predictions are obtained is not clearly explained in the corrosion prediction studies.

In this work, we applied different models (ANN, RF, AdaBoost, GBRT, and LightGBM) for regression to predict the **dmax** of oil and gas pipelines. Then the best models were identified and further optimized. More importantly, this research aims to explain the black box nature of ML in predicting corrosion in response to the previous research gaps. The study visualized the final tree model, explained how some specific predictions are obtained using SHAP, and analyzed the global and local behavior of the model in detail. Moreover, ALE plots were utilized to describe the main and interaction effects of features on predicted results. This study emphasized that interpretable ML does not sacrifice accuracy or complexity inherently, but rather enhances model predictions by providing human-understandable interpretations and even helps discover new mechanisms of corrosion.

RESULTS AND DISCUSSION

Data analysis and pre-processing

Variance, skewness, kurtosis, and coefficient of variation are used to describe the distribution of a set of data, and these metrics for the quantitative variables in the data set are shown in Table 1. Specifically, the kurtosis and skewness indicate the difference from the normal distribution. The coefficient of variation (CV) indicates the likelihood of the outliers in the data. It is generally considered that outliers are more likely to exist if the CV is higher than 0.15. As shown in Table 1, the CV for all variables exceed 0.15 excluding **pp** (pipe/soil potential) and **bd** (bulk density), which means that outliers may exist in the applied dataset. In addition, the variance, kurtosis, and skewness of most the variables are large, which further increases this possibility.

The violin plot reflects the overall distribution of the original data. Box plots are used to quantitatively observe the distribution of the data, which is described by statistics such as the median, 25% quantile, 75% quantile, upper bound, and lower bound. The box contains most of the normal data, while those outside the upper and lower boundaries of the box are the potential outliers. Figure 1 shows the combination of the violin plots and box plots applied to the quantitative variables in the database. It can be found that there are potential outliers in all features (variables) except **rp** (redox potential). Considering the actual meaning of the features and the scope of the theory, we found 19 outliers, which are more than the outliers marked in the original database, and removed them.

Although the coating type in the original database is considered as a discreet sequential variable and its value is assigned according to the scoring model[30], the process is very complicated. To make the categorical variables suitable for ML regression models, one-hot encoding was employed. Table 2 shows the one-hot encoding of the coating type and soil type. For example, each soil type is represented by a 6-bit status register, where clay and clay loam are coded as 100000 and 010000, respectively. The status register bits are named as Class_C, Class_CL, Class_SC, Class_SCL, Class_SL, and Class_SYCL accordingly. The one-hot encoding also implies an increase in feature dimension, which will be further filtered in the later discussion.

Table 1. Description of the overall distribution of the variables.

Variable	Variance	Kurtosis	Skewness	CV
dmax	4.18	7.07	2.52	1.011
t	83.14	0.58	0.72	0.397
pH	0.86	0.52	0.51	0.152
pp	0.06	3.90	−1.18	−0.273
re	3127.03	8.73	2.53	1.115
wc	44.35	7.43	1.51	0.279
bd	0.008	−0.18	0.12	0.068
cc	5648.94	22.42	3.99	1.575
bc	641.6	17.45	3.76	1.288
sc	28285.2	15.73	3.33	1.099
rp	7307.5	−0.86	0.23	0.512

Performance evaluation of the models

This section covers the evaluation of models based on four different EL methods (RF, AdaBoost, GBRT, and LightGBM) as well as the ANN framework. It is noted that the ANN structure involved in this study is the BPNN with only one hidden layer. Five statistical indicators, mean absolute error (MAE), coefficient of determination (R2), mean square error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE) were used to evaluate and compare the validity and accuracy of the prediction results for 40 test samples.

Table 3 reports the average performance indicators for ten replicated experiments, which indicates that the EL models provide more accurate predictions for the **dmax** in oil and gas pipelines compared to the ANN model. In general, the superiority of ANN is learning the information from the complex and high-volume data, but tree models tend to perform better with smaller dataset. The pre-processed dataset in this study contains 240 samples with 21 features, and the tree model is more superior at handing this data volume. The results show that RF, AdaBoost, GBRT, and LightGBM are all tree models that outperform ANN on the studied dataset.

As shown in Fig. 2a, the prediction results of the AdaBoost model fit the true values best under the condition that all models use the default parameters. The scatters of the predicted versus true values are located near the perfect line as in Fig. 2b. Further analysis of the results in Table 3 shows that the Adaboost model is

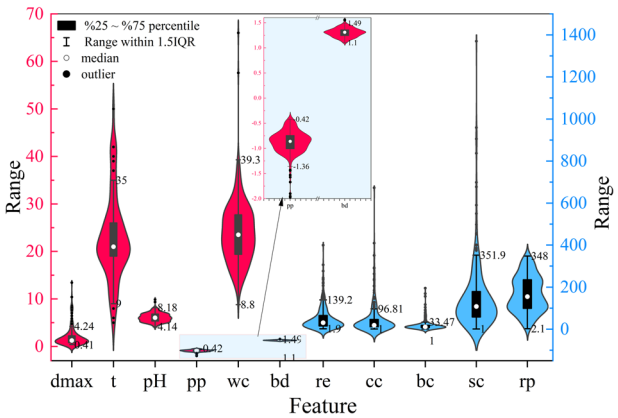

Fig. 1 Data distribution and outliers. A combination of box plot and violin plot to visualize the distribution and outliers of the data.

Table 2. One-hot encoding of coating and soil type.

Variable		One-hot code					
		Class_C	Class_CL	Class_SC	Class_SCL	Class_SL	Class_SYCL
Class	C	1	0	0	0	0	0
	CL	0	1	0	0	0	0
	SC	0	0	1	0	0	0
	SCL	0	0	0	1	0	0
	SL	0	0	0	0	1	0
	SYCL	0	0	0	0	0	1
		ct_AEC	ct_CTC	ct_FBE	ct_NC	ct_WTC	
ct	AEC	1	0	0	0	0	
	CTC	0	1	0	0	0	
	FBE	0	0	1	0	0	
	NC	0	0	0	1	0	
	WTC	0	0	0	0	1	

Table 3. Average performance of models.

Average Evaluation Metrics	Model				
	RF	AdaBoost	GBRT	LightGBM	ANN
MSE	0.892	0.406	0.669	0.716	1.176
RMSE	0.934	0.624	0.799	0.814	1.084
MAE	0.639	0.430	0.552	0.576	0.792
R^2	0.764	0.895	0.830	0.804	0.733
MAPE	30.585	25.772	31.293	30.843	37.278

superior to the other models in all metrics among EL, with R^2 and RMSE values of 0.895 and 0.624, respectively. The reason is that AdaBoost, which runs sequentially, enables to give more attention to the missplitting data and constantly improve the model, making the sequential model more accurate than the simple parallel model.

Adaboost model optimization

The AdaBoost was identified as the best model in the previous section. The screening of features is necessary to improve the performance of the Adaboost model. A preliminary screening of these features is performed using the AdaBoost model to calculate the importance of each feature on the training set via "feature_importances_" function built into the Scikit-learn python module. As shown in Fig. 3, pp has the strongest contribution with an importance above 30%, which indicates that this feature is extremely important for the **dmax** of the pipeline. The **pp** (protection potential, natural potential, E_{on} or E_{off} potential) is a parameter related to the size of the electrochemical half-cell and is an indirect parameter of the surface state of the pipe at a single location, which covers the macroscopic conditions during the assessment of the field conditions[31]. The industry generally considers steel pipes to be well protected at **pp** below −850 mV[32].

pH and **cc** (chloride content) are another two important environmental factors, with importance of 15.9% and 12.3%, respectively. The acidity and erosion of the soil environment are enhanced at lower pH, especially when it is below 5[1]. Chloride ions are a key factor in the depassivation of naturally occurring passive film. At concentration thresholds, chloride ions decompose this passive film under microscopic conditions, accelerating corrosion at specific locations[33]. The service time of the pipeline is also an important factor affecting the **dmax**, which is in line with basic fundamental experience and intuition. The contribution of all the above four features exceeds 10%, and the cumulative contribution exceeds 70%, which can be largely regarded as key features. Meanwhile, the calculated results of the importance of Class_SC, Class_SL, Class_SYCL, ct_AEC, and ct_FBE are equal to 0, and thus they are removed from the selection of key features. It is worth noting that this does not absolutely imply that these features are completely independent of the **damx**. Probably due to the small sample in the dataset, the model did not learn enough information from this dataset.

In order to identify key features, the correlation between different features must be considered as well, because strongly related features may contain the redundant information. They may obscure the relationship between the **dmax** and features, and reduce the accuracy of the model[34]. Figure 4 reports the matrix of the Spearman correlation coefficients between the different features, which is used as a metric to determine the related strength between these features. **bd** (soil bulk density) and class_SCL are closely correlated with the coefficient above 0.75, and t shows a correlation of 0.78 with ct_CTC (coal-tar-coated coating). Correlation coefficient 0.6–0.8 can be considered as strongly correlated. Based on the data characteristics and calculation results of this study, we used the median 0.7 as the threshold value. Strongly correlated (>0.7) features imply the similarity in nature, and thus the feature dimension can be reduced by removing less important factors from the strongly correlated features. As an example, the correlation coefficients of **bd** with Class_C (clay) and Class_SCL (sandy clay loam) are −0.80 and 0.75, respectively, which indicates a close monotonic relationship between **bd** and these two features. Specifically, class_SCL implies a higher **bd**, while Claa_C is the contrary. As determined by the AdaBoost model, **bd** is more important than the other two factors, and thus so Class_C and Class_SCL are considered as the redundant features and removed from the selection of key features. Only **bd** is considered in the final model, essentially because it implys the Class_C and Class_SCL. Similarly, ct_WTC and ct_CTC are considered as redundant.

Table 4 summarizes the 12 key features of the final screening. Among soil and coating types, only Class_CL and ct_NC are considered. In addition, the association of these features with the **dmax** are calculated and ranked in Table 4 using GRA, and they all exceed 0.9, verifying that these features are crucial. The accuracy of the AdaBoost model with these 12 key features as input is maintained ($R^2 = 0.96$) and the model is more robust.

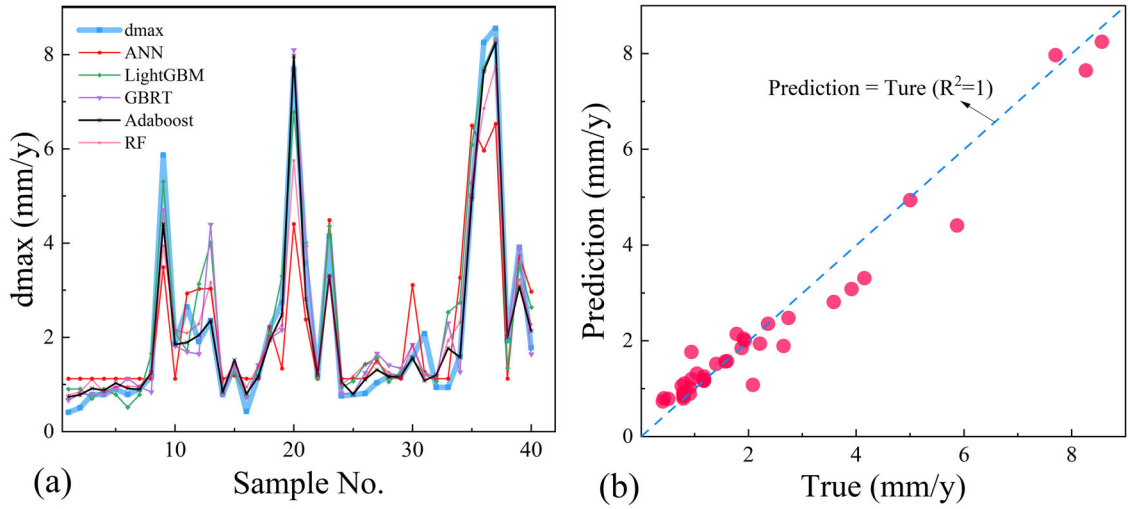

Fig. 2 Prediction results of models. a Predictions of different models on the test set, **b** scatter plot of predictions using Adaboost model.

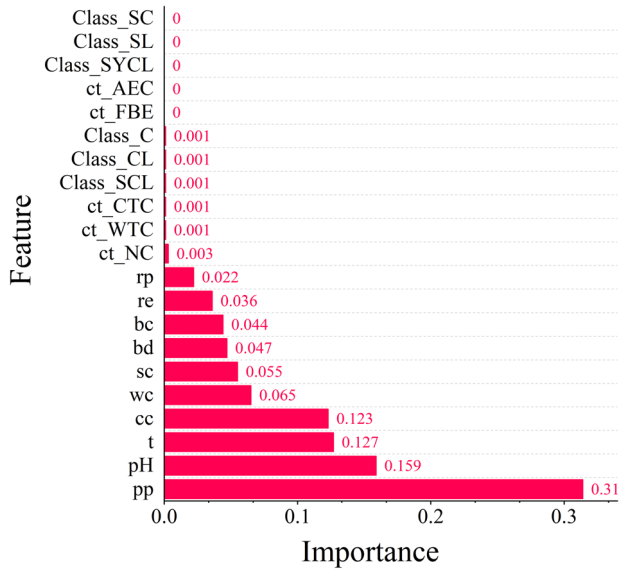

Fig. 3 The relative importance of 21 features according to AdaBoost. The important features will be used for the subsequent optimization step.

There are numerous hyperparameters that affect the performance of the AdaBoost model, including the type and number of base estimators, loss function, learning rate, etc. In this study, the base estimator is set as decision tree, and thus the hyperparameters in the decision tree are also critical, such as the maximum depth of the decision tree (max_depth), the minimum sample size of the leaf nodes, etc. Figure 5 shows how the changes in the number of estimators and the max_depth affect the performance of the AdaBoost model with the experimental dataset. The line indicates the average result of 10 tests, and the color block is the error range.

It can be found that as the estimator increases (other parameters are default, learning rate is 1, number of estimators is 50, and the loss function is linear), the MSE and MAPE of the model decrease, while R^2 increases. The model performance reaches a better level and is maintained when the number of estimators exceeds 50. In addition, the error bars of the model also decrease gradually with the increase of the estimators, which means that the model is more robust. The max_depth significantly affects the performance of the model. The overall performance is improved as the increase of the max_depth. However, once the max_depth exceeds 5, the model tends to be stable with the R^2, MSE, and MAEP equal to 0.950, 0.225, and 0.302, accordingly.

To further determine the optimal combination of hyperparameters, Grid Search with Cross Validation strategy is used to search for the critical parameters. In this study, only the max_depth is considered in the hyperparameters of the decision tree due to the small sample size. The candidate for the number of estimator is set as: [10, 20, 50, 100, 150, 200, 250, 300]. The candidates for the loss function, the max_depth, and the learning rate are set as ['linear', 'square', 'exponential'], [3, 5, 7, 9, 12, 15, 18, 21, 25], and [0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1], accordingly. The total search space size is 8×3×9×7. Finally, the best candidates for the max_depth, loss function, learning rate, and number of estimators are 12, 'liner', 0.1, and 50, accordingly.

Model visualization and interpretation

Previous ML prediction models usually failed to clearly explain how these predictions were obtained, and the same is true in corrosion prediction, which made the models difficult to understand. To explore how the different features affect the prediction overall is the primary task to understand a model. In this work, SHAP is used to interpret the prediction of the AdaBoost model on the entire dataset, and its values are used to quantify the impact of features on the model output. A negative SHAP value means that the feature has a negative impact on the prediction, resulting in a lower value for the model output. Conversely, a positive SHAP value indicates a positive impact that is more likely to cause a higher **dmax**.

Figure 6a depicts the global distribution of SHAP values for all samples of the key features, and the colors indicate the values of the features, which have been scaled to the same range. Blue and red indicate lower and higher values of features. As can be seen that pH has a significant effect on the **dmax**, and lower pH usually shows a positive SHAP, which indicates that lower pH is more likely to improve **dmax**. Conversely, a higher pH will reduce the **dmax**. In addition, previous studies showed that the corrosion rate on the outside surface of the pipe is higher when the concentration of chloride ions in the soil is higher, and the deeper pitting corrosion produced[35]. This is consistent with the depiction of feature cc in Fig. 6a, where higher values of **cc** (chloride content) have a reasonably positive effect on the **dmax** of the pipe, while lower values have negative effect. Similarly, higher **pp** (pipe/soil potential) significantly increases the

Fig. 4 Matrix of the Spearman correlation coefficients between the different features. Strongly correlated (>0.7) features indicate the presence of redundant variables, which guide the screening of features.

Table 4. The result of gray relational analysis.

Gray relational analysis					
Feature	Association	Rank	Feature	Association	Rank
t	0.977	1	**rp**	0.968	7
wc	0.976	2	**sc**	0.967	8
bd	0.974	3	**bc**	0.965	9
pH	0.973	4	**re**	0.96	10
cc	0.972	5	Class_CL	0.943	11
pp	0.972	6	ct_NC	0.940	12

probability of larger pitting depth, while lower **pp** reduces the **dmax**. This is because sufficiently low **pp** is required to provide effective protection to the pipeline. **t** (pipeline age) and **wc** (water content) have the similar effect on the **dmax**, and higher values of features show positive effect on the **dmax**, which is completely opposite to the effect of **re** (resistivity). The remaining features such as ct_NC and **bc** (bicarbonate content) present less effect on the pitting globally.

The average SHAP values are also used to describe the importance of the features. As shown in Fig. 6b, cc has the highest importance with an average absolute SHAP value of 0.54. It means that the **cc** of all samples in the AdaBoost model improves the **dmax** by 0.54 mm on average. The next is pH, which has an average SHAP value of 0.48. **pp** and **t** are the other two main features with SHAP values of 0.27 and 0.3, respectively. Globally, **cc**, pH, **pp**, and **t** are the four most important features affecting the **dmax**, which is generally consistent with the results discussed in the previous section.

Although the overall analysis of the AdaBoost model has been done above and revealed the macroscopic impact of those features on the model, the model is still a black box. Visualization and local interpretation of the model can open up the black box to help us understand the mechanism of the model and explain the interactions between features. As previously mentioned, the AdaBoost model is computed sequentially from multiple decision

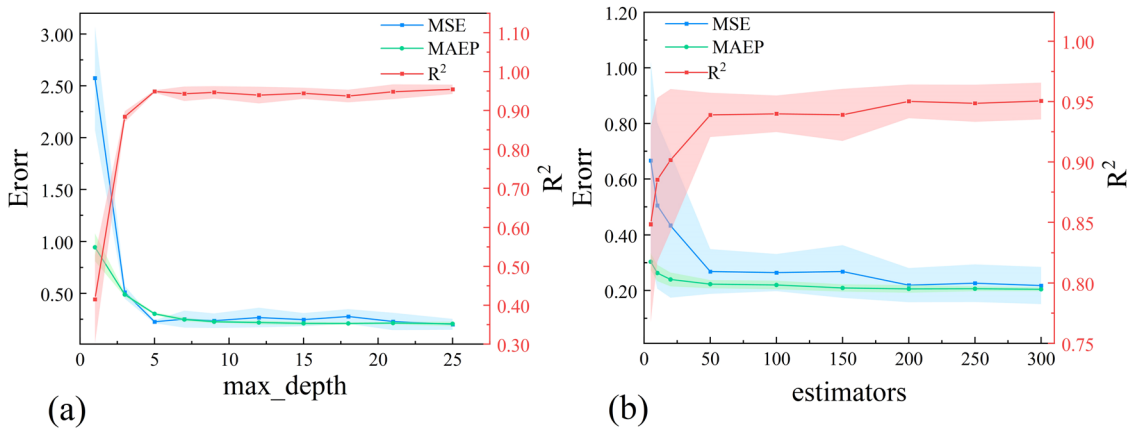

Fig. 5 Performance of the model with different hyperparameter values. a Variation of AdaBoost model performance with maximum depth of decision tree, **b** Variation of AdaBoost model performance with the number of base estimators.

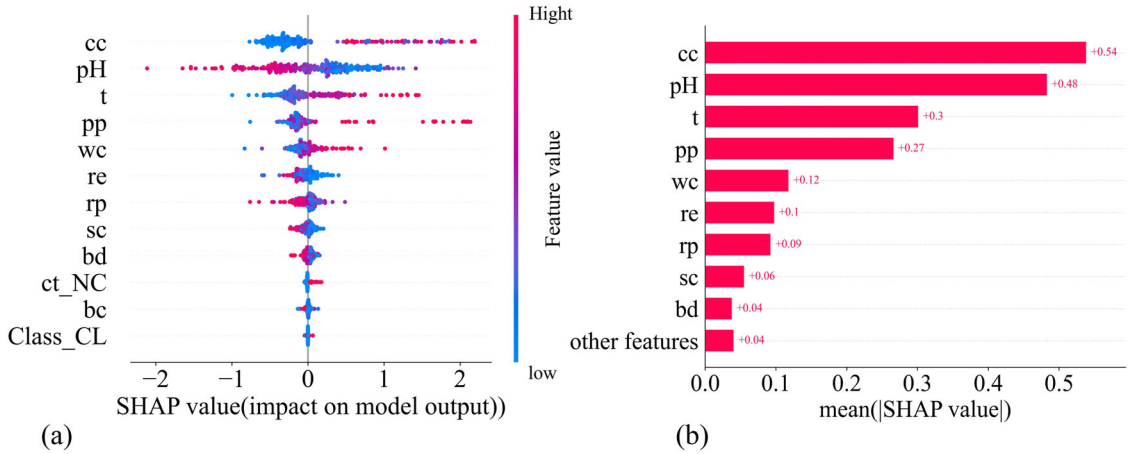

Fig. 6 SHAP global interpretation. a SHAP value for all samples of key features, **b** Mean SHAP value of features.

trees, and we creatively visualize the final decision tree. According to the optimal parameters, the max_depth (maximum depth) of the decision tree is 12 layers. Figure 7 shows the first 6 layers of this decision tree and the traces of the growth (prediction) process of a record. Taking the first layer as an example, if a sample has a **pp** value higher than −0.60 V, then it will grow along the right subtree, otherwise it will turn to the left subtree. The sample tracked in Fig. 7 is branched five times and the prediction is locked at 0.97 after discriminating the values of **pp**, **cc**, pH, and **t**. It should be noted that this is the result of the calculation after 5 layer of decision trees, and the result after the full decision tree is 0.71, which is very close to the actual result. This decision tree is the basis for the model to make predictions. Once the values of these features are measured in the applicable environment, we can follow the graph and get the **dmax**.

In addition to the global interpretation, Fig. 8 shows the instances of local interpretations (particular prediction) obtained from SHAP values. The gray vertical line in the middle of the SHAP decision plot (Fig. 8a) marks the base value of the model, and the colored ones are the prediction lines, which show how the model accumulates from the base value to the final outputs starting from the bottom of the plots. Figure 8a shows the prediction lines for ten samples numbered 140–150, in which the more upper features have higher influence on the predicted results. It is consistent with the importance of the features. Further, pH and cc demonstrate the opposite effects on the predicted values of the model for the most part. Figure 8b shows the SHAP waterfall plot for sample numbered 142 (black dotted line in Fig. 8a), which interprets the unique contribution of the variables to the result at any given point. In this plot, $E[f(x)] = 1.9$ is the baseline (average expected value) and the final value is $f(x) = 1.57$, which is also the predicted value for this instance. The SHAP value in each row represents the contribution and interaction of this feature to the final predicted value of this instance. For example, the pH of 5.56 has a positive effect on the **damx**, which adds 0.32 to the prediction from the baseline. However, **cc** (14.4 ppm) has a negative effect on the **damx**, which decreases the predicted result by 0.28. It indicates that the content of chloride ions, 14.4 ppm, has not yet reached the threshold to promote pitting. Interestingly, the **rp** of 328 mV in this instance shows a large effect on the results, but **t** (19 years) does not. However, none of these showed up in the global interpretation, so further quantification of the impact of these features on the predicted results is requested. Figure 8c shows this SHAP force plot, which can be considered as a horizontal projection of the waterfall plot and clusters the features that push the prediction higher (red) and lower (blue).

To further depict how individual features affect the model's predictions continuously, ALE main effect plots are employed. Figure 9 shows the ALE main effect plots for the nine features with significant trends. The ranking over the span of ALE values for these features is generally consistent with the ranking of feature importance discussed in the global interpretation, which indirectly validates the reliability of the ALE results.

As shown in Fig. 9a, the ALE values of the **dmax** present a monotonically increasing relationship with the **cc** in the overall. That is, the higher the amount of chloride in the environment, the

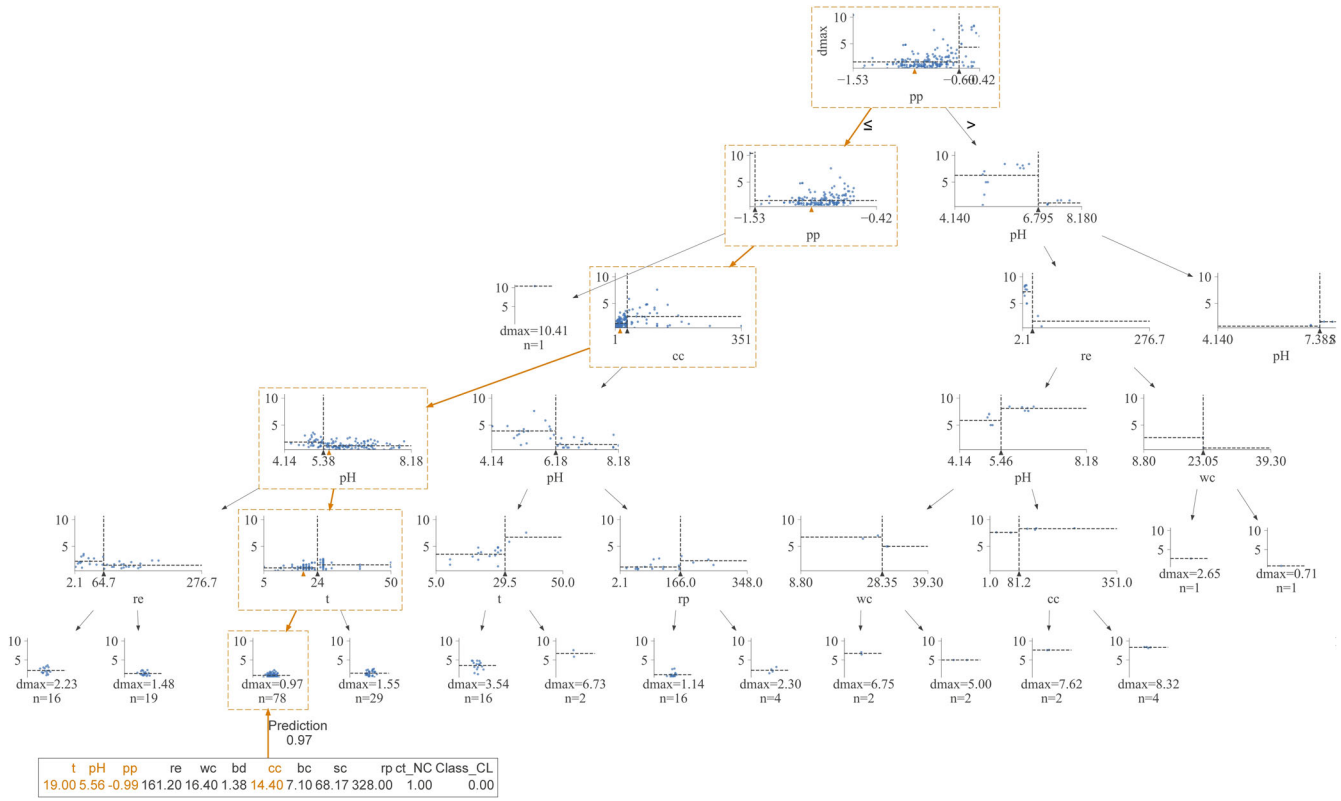

Fig. 7 Visualization of decision trees. The splitting logic for the first six layers of the part of the decision tree generated by AdaBoost, and the variables that have been calculated are marked in yellow in the bottom box.

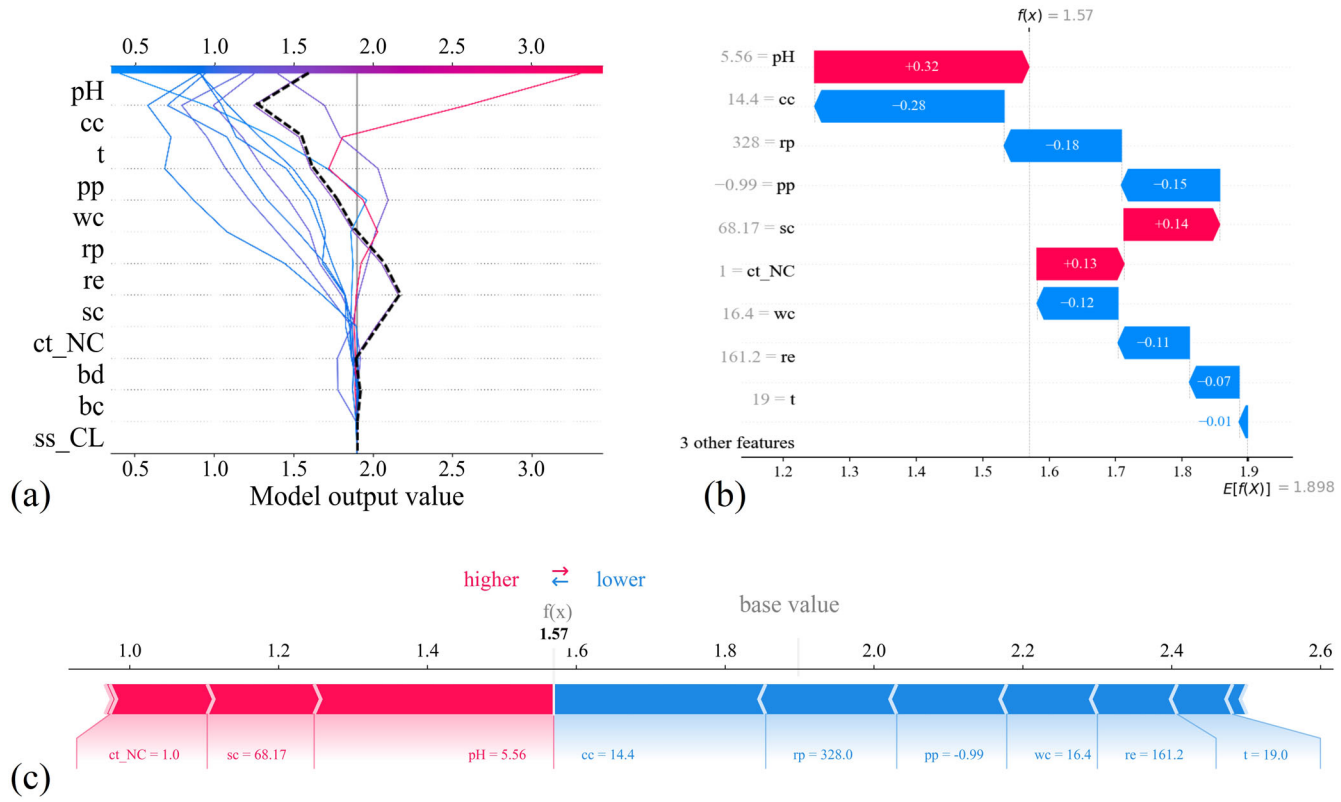

Fig. 8 SHAP local interpretation. a SHAP prediction lines for ten instances, **b** SHAP waterfall plot on selected instance, **c** SHAP force plot on selected instance.

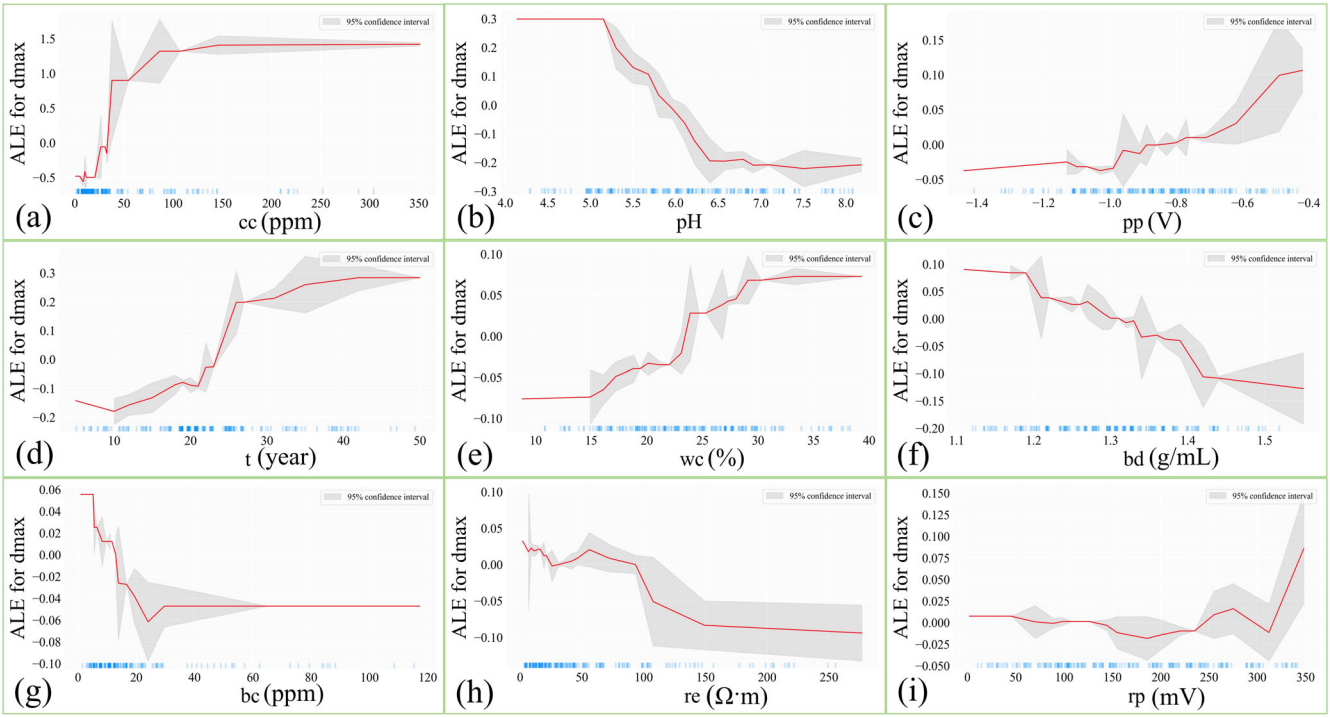

Fig. 9 ALE main effect plot for dmax. ALE versus **a cc**, **b** pH, **c pp**, **d t**, **e wc**, **f bd**, **g bc**, **h re**, **i rp**.

larger the **dmax**. It is interesting to note that **dmax** exhibits a very strong sensitivity to **cc** (chloride content), and the ALE value increases sharply as **cc** exceeds 20 ppm. It means that the pipeline will obtain a larger **dmax** owing to the promotion of pitting by chloride above the critical level. However, the excitation effect of chloride will reach stability when the **cc** exceeds 150 ppm, and chloride are no longer a critical factor affecting the **dmax**. The reason is that high concentration of chloride ions cause more intense pitting on the steel surface, and the developing pits are covered by massive corrosion products, which inhibits the development of the pits[36]. In addition, there is not a strict form of the corrosion boundary in the complex soil environment, the local corrosion will be more easily extended to the continuous area under higher chloride content, which results in a corrosion surface similar to the general corrosion and the corrosion pits are erased[35]. pH is a local parameter that modifies the surface activity mechanism of the environment surrounding the pipe. Low pH environment lead to active corrosion and may create local conditions that favor the corrosion mechanism of sulfate-reducing bacteria[31]. The corrosion rate increases as the pH of the soil decreases in the range of 4–8.5, and the **dmax** is larger, as shown in Fig. 9b. The ALE values of **dmax** are monotonically increasing with both **t** and **pp** (pipe/soil potential), as shown in Fig. 9c and d. It means that the longer the exposure time of pipelines, the more positive potential of the pipe/soil is, and then the larger pitting depth is more accessible. From Fig. 9c, it is further found that the **dmax** increases rapidly for the values of **pp** above −0.8 V, while the pipeline is well protected for values below −0.8 V.

wc (water content) is also key to inducing external corrosion in oil and gas pipelines, and this parameter depends on physical factors such as soil skeleton, pore structure, and density[31]. As the **wc** increases, the corrosion rate of metals in the soil increases until reaching a critical level. Then, with the further increase of the **wc**, the oxygen supply to the metal surface decreases and the corrosion rate begins to decrease[37]. The critical **wc** is related to the soil type and its characteristics, the type of pipe steel, the exposure conditions of the metal, and the time of the soil exposure. The curve in Fig. 9e depicts a positive correlation between **dmax** and **wc** within 35%, but it is not able to determine the critical **wc**, which could be explained by the fact that the sample of the data set is still not extensive enough. The high **wc** of the soil also leads to the growth of corrosion-inducing bacteria in contact with buried pipes, which may increase pitting[38]. With the increase of **bd** (bulk density), **bc** (bicarbonate content), and **re** (resistivity), **dmax** presents a decreasing trend, and all of them are strongly sensitive within a certain range. Once **bc** is over 20 ppm or **re** exceeds 150 Ω·m, **damx** remains stable, as shown in Fig. 9f, g, h. **rp** (redox potential) has no significant effect on **dmax** in the range of 0–300 mV, but the oxidation capacity of the soil is enhanced and pipe corrosion is accelerated at higher **rp**[39].

In addition to the main effect of single factor, the corrosion of the pipeline is also subject to the interaction of multiple factors. The interactio n effect of the two features (factors) is known as the second-order interaction. The ALE second-order interaction effect plot indicates the additional interaction effects of the two features without including their main effects. The red and blue represent the above and below average predictions, respectively. Figure 10a shows the ALE second-order interaction effect plot for pH and **pp**, which reflects the second-order effect of these features on the **dmax**. For low pH and high **pp** (zone A) environments, an additional positive effect on the prediction of **dmax** is seen. High pH and high **pp** (zone B) have an additional negative effect on the prediction of **dmax**. However, low pH and **pp** (zone C) also have an additional negative effect. As shown in Fig. 10b, Pourbaix diagram of the Fe-H_2O system illustrates the main areas of immunity, corrosion, and passivation condition over a wide range of pH and potential. **pp** is the potential of the buried pipeline relative to the Cu/$CuSO_4$ electrode, which is the free corrosion potential (E_{corr}) of the pipeline[40]. While the potential in the Pourbaix diagram is the potential of Fe relative to the standard hydrogen electrode E_{corr} in water. Despite the difference in potential, the Pourbaix diagram can still provide a valid guide for the protection of the pipeline. It is generally considered that the cathodic protection of pipelines is favorable if the pp is below −0.85 V[40,41]. In Fig. 10, zone A is not within the protection

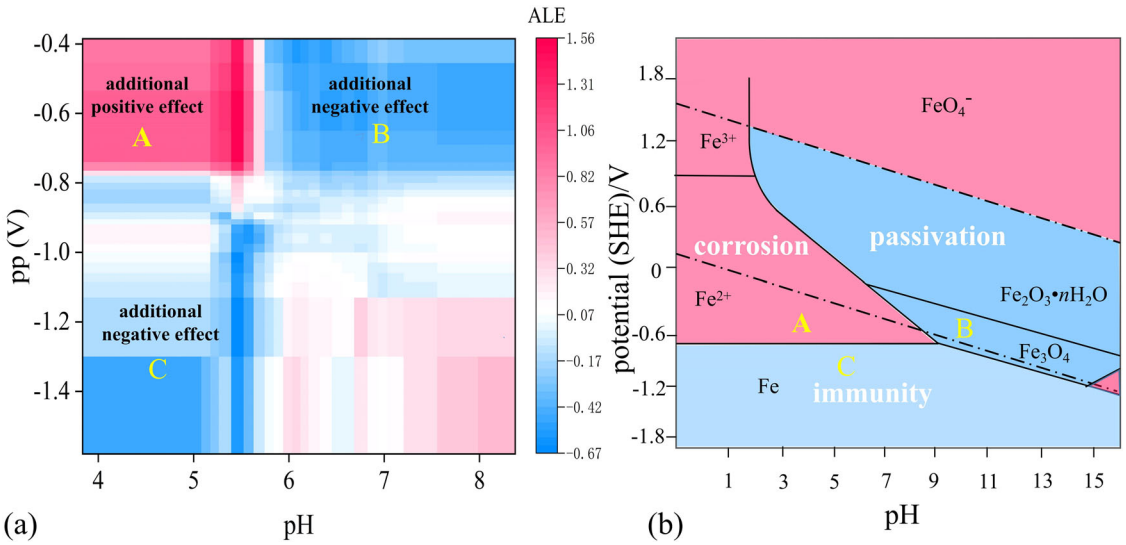

Fig. 10 Interaction of pH and pp. a ALE second-order interaction effect plot for pH and **pp**, **b** Pourbaix diagram of the Fe-H_2O.

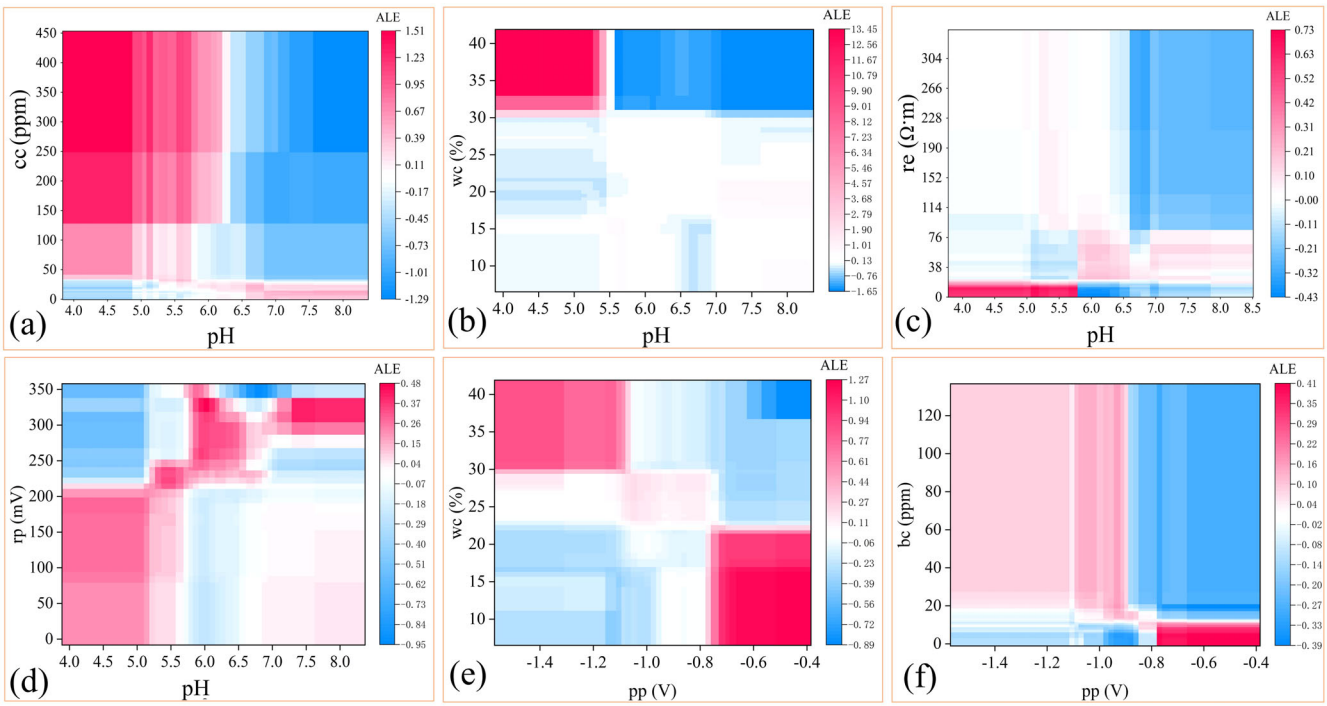

Fig. 11 ALE second-order interaction plots. ALE plots for **a** pH and **cc**, **b** pH and **wc**, **c** pH and **re**, **d** pH and **rp**, **e pp** and **wc**, **f pp** and **bd**.

potential and corresponds to the corrosion zone of the Pourbaix diagram, where the pipeline has a severe tendency to corrode, resulting in an additional positive effect on **dmax**. Zones B and C correspond to the passivation and immunity zones, respectively, where the pipeline is well protected, resulting in an additional negative effect. In general, the calculated ALE interaction effects are consistent with the corrosion experience.

Similarly, more interaction effects between features are evaluated and shown in Fig. 11. Figure 11a reveals the interaction effect between pH and **cc**, showing an additional positive effect on the **dmax** for the environment with low pH and high **cc**. Although the increase of **dmax** with increasing **cc** was demonstrated in the previous analysis, high pH and **cc** show an additional negative effect on the prediction of the **dmax**, which implies that high pH reduces the promotion of corrosion caused by chloride. The interaction of low pH and high **wc** has an additional positive effect on **dmax**, as shown in Fig. 11b. That is, lower pH amplifies the effect of **wc**. Basic and acidic soils may have associated corrosion, depending on the resistivity[1,42]. According to the standard BS EN 12501-2:2003, Amaya-Gomez et al.[42] reported a corrosion classification diagram for combined soil resistivity and pH, which indicates that oil and gas pipelines in low soil resistivity are more susceptible to external corrosion at low pH. This is verified by the interaction of pH and **re** depicted in Fig. 11c, where low pH and **re** additionally contribute to the **dmax**. In addition, low pH and low **rp** give an additional promotion to the **dmax**, while high pH and **rp** give an additional negative effect as shown in Fig. 11d. In the previous discussion, it has been pointed out that

the corrosion tendency of the pipelines increases with the increase of **pp** and **wc**. As shown in Fig. 11e, this law is still reflected in the second-order effects of **pp** and **wc**. Within the protection potential, the increasing of **wc** leads to an additional positive effect, i.e., the pipeline corrosion is further promoted. In the lower **wc** environment, the high **pp** causes an additional negative effect, as the high potential increases the corrosion tendency of the pipelines. The difference is that high **pp** and high **wc** produce additional negative effects, which may be attributed to the formation of corrosion product films under severe corrosion, and thus corrosion is depressed. In addition, Fig. 11f indicates that the effect of **bc** on dmax is further amplified at high **pp** condition. More second-order interaction effect plots between features will be provided in Supplementary Figures.

In the above discussion, we analyzed the main and second-order interactions of some key features, which explain how these features in the model affect the prediction of **dmax**. However, the effect of third- and higher-order effects of the features on **dmax** were done discussed, since high order effects are difficult to interpret and are usually not as dominant as the main and second order effects[43].

In summary, five valid ML models were used to predict the maximum pitting depth (**damx**) of the external corrosion of oil and gas pipelines using realistic and reliable monitoring data sets. Spearman correlation coefficient, GRA, and AdaBoost methods were used to evaluate the importance of features, and the key features were screened and an optimized AdaBoost model was constructed. In addition, This paper innovatively introduces interpretability into corrosion prediction. The interpretations and transparency frameworks help to understand and discover how environment features affect corrosion, and provide engineers with a convenient tool for predicting **dmax**. The main conclusions are summarized below.

(1) Compared with ANN, RF, GBRT, and lightGBM, AdaBoost can predict the **dmax** of the pipeline more accurately, and its performance index R^2 value exceeds 0.95 after optimization.
(2) **cc** (chloride content), pH, **pp** (pipe/soil potential), and **t** (pipeline age) are the four most important factors affecting **dmax** in several evaluation methods. While coating and soil type show very little effect on the prediction in the studied dataset.
(3) The ALE values of **dmax** present the monotonic increase with increasing **cc**, **t**, **wc** (water content), **pp**, and **rp** (redox potential), which indicates that the increase of **cc**, **wc**, **pp**, and **rp** in the environment all contribute to the **dmax** of the pipeline. Conversely, increase in pH, **bd** (bulk density), **bc** (bicarbonate content), and **re** (resistivity) reduce the **dmax**.
(4) The interaction of features shows a significant effect on **dmax**. pH exhibits second-order interaction effects on **dmax** with **pp**, **cc**, **wc**, **re**, and **rp**, accordingly. At the extreme values of the features, the interaction of the features tends to show the additional positive or negative effects.

METHODS

Collection and description of experimental data

Sufficient and valid data is the basis for the construction of artificial intelligence models. The establishment and sharing practice of reliable and accurate databases is an important part of the development of materials science under the new paradigm of materials science development. The experimental data for this study were obtained from the database of Velázquez et al.[30], which covers various important parameters in the initiation and growth of corrosion defects. This database contains 259 samples of soil and pipe variables for an onshore buried pipeline that has been in operation for 50 years in southern Mexico. These environmental variables include soil resistivity, pH, water content, redox potential, bulk density, and concentration of dissolved chloride, bicarbonate and sulfate ions, and pipe/soil potential. The service time of the pipe, the type of coating, and the soil are also covered. Soil samples were classified into six categories: clay (C), clay loam (CL), sandy loam (SCL), and silty clay (SC) and silty loam (SL), silty clay loam (SYCL), based on the relative proportions of sand, silty sand, and clay. Coating types include noncoated (NC), asphalt-enamel-coated (AEC), wrap-tape-coated (WTC), coal-tar-coated (CTC), and fusion-bonded-epoxy-coated (FBE). The maximum pitting depth (**dmax**), defined as the maximum depth of corrosive metal loss for diameters less than twice the thickness of the pipe wall, was measured at each exposed pipeline segment. Nine outliers had been pointed out by simple outlier observations, and the complete dataset is available in the literature[30] and a brief description of these variables is given in Table 5. Figure 12 shows the distribution of the data under different soil types.

Data pre-processing

Data pre-processing is a necessary part of ML. In this study, we mainly consider outlier exclusion and data encoding in this session. Variance, skewness, kurtosis, and CV are used to profile the global distribution of the data. Specifically, Skewness describes the symmetry of the distribution of the variable values, Kurtosis describes the steepness, Variance describes the dispersion of the data, and CV combines the mean and standard deviation to reflect the degree of data variation. These statistical values can help to determine if there are outliers in the dataset. If the CV is greater than 15%, there may be outliers in this dataset. Combining the kurtosis and skewness values we can further analyze this possibility. To further identify outliers in the dataset, the interquartile range (IQR) is commonly used to determine the boundaries of outliers. The first quartile (25% quartile) is Q1 and the third quartile (75% quartile) is Q3, then IQR = Q3-Q1. Specifically, for samples smaller than Q1-1.5IQR (lower bound), and larger than Q3 + 1.5IQR (upper bound) are considered outliers and should be excluded.

Table 5. Basic descriptive statistics of variables.

Variable	Symbol	Units	Max	Min	Mean	Std	Med
Maximum pitting depth	**dmax**	mm	13.44	0.41	2.024	2.05	1.25
Pipeline age	**t**	year	50	5	22.98	9.12	21
pH	**pH**	—	9.88	4.14	6.13	0.93	6.05
Pipe/soil potential	**pp**	V	−0.42	−1.97	−0.87	0.24	−0.86
Resistivity	**re**	Ω · m	399.5	1.9	50.14	55.92	30.9
Water content	**wc**	%	66	8.8	23.89	6.66	23.5
Bulk density	**bd**	g/mL	1.56	1.1	1.30	0.09	1.31
Chloride content	**cc**	ppm	672.7	1	47.72	75.16	19.02
Bicarbonate content	**bc**	ppm	195.2	1	19.67	25.33	11.1
Sulfate content	**sc**	ppm	1370.2	1	152.97	168.18	108.4
Redox potential	**rp**	mV	348	2.1	167.04	85.48	155
Coating type	**ct**	—	—	—	—	—	—
Soil classification	**Class**	—	—	—	—	—	—

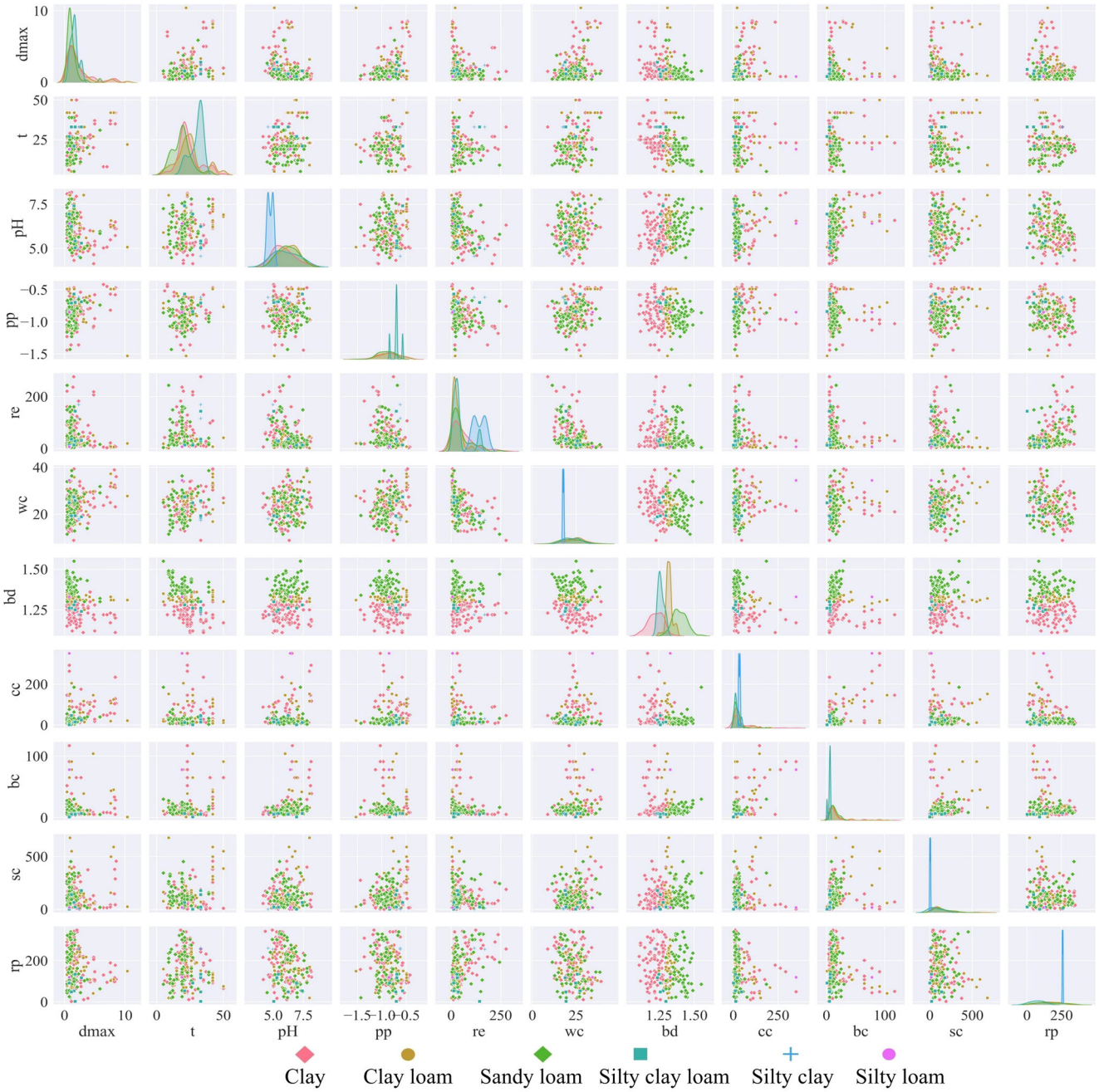

Fig. 12 Data distribution under different soil types. Samples were classified by soil type, and the overall distribution of these samples on the two feature dimensions was observed. Plots on the diagonal depicting the distribution of individual features for different soil type samples.

The one-hot encoding can represent categorical data well and is extremely easy to implement without complex computations. The approach is to encode different classes of classification features using status registers, where each class has its own independent bits and only one of them is valid at any given time. That is, only one bit is 1 and the rest are zero.

Ensemble learning

Ensemble learning (EL) is an algorithm that combines many base machine learners (estimators) into an optimal one to reduce error, enhance generalization, and improve model prediction[44]. EL with decision tree based estimators is widely used. Generally, EL can be classified into parallel and serial EL based on the way of combination of base estimators. Parallel EL models, such as the classical Random Forest (RF), use bagging to train decision trees independently in parallel, and the final output is an average result. Sequential EL reduces variance and bias by creating a weak predictive model and iterating continuously using boosting techniques. This is true for AdaBoost, gradient boosting regression tree (GBRT) and light gradient boosting machine (LightGBM) models. We selected four potential algorithms from a number of EL algorithms by considering the volume of data, the properties of the algorithms, and the results of pre-experiments. The following part briefly describes the mathematical framework of the four EL models.

RF is a strongly supervised EL method that consists of a large number of individual decision trees that operate as a whole. Each individual tree makes a prediction or classification, and the

prediction or classification with the most votes becomes the result of the RF[45]. The method consists of two phases to achieve the final output. In the first stage, RF uses bootstrap aggregating approach to select input features randomly and training datasets to build multiple decision trees. This random property reduces the correlation between individual trees, and thus reduces the risk of over-fitting. In the second stage, the average result of the predictions obtained from the individual decision tree is calculated as follow[25]:

$$y(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x) \tag{1}$$

Where, y_i represents the i-th decision tree, and the total number of trees is n. y is the target output, and x denotes the feature vector of the input.

AdaBoost is a powerful iterative EL technique that creates a powerful predictive model by merging multiple weak learning models[46]. The general form of AdaBoost is as follow:

$$F_T(X) = \sum_{t=1}^{T} f_t(X) \tag{2}$$

Where f_t denotes the weak learner and X denotes the feature vector of the input. Each iteration generates a new learner using the training dataset to evaluate all samples. During the process, the weights of the incorrectly predicted samples are increased, while the correct ones are decreased. Meanwhile, a new hypothetical weak learner will be added in each iteration to minimize the total training error, as follow.

$$E_t = \sum\nolimits_i E[F_{t-1}(X_i) + a_t h(X_i)] \tag{3}$$

$F_{t\text{-}1}$ denotes the weak learner obtained from the previous iteration, and $f_t(X) = a_t h(X)$ is the improved weak learner. Eventually, AdaBoost forms a single strong learner by combining several weak learners.

Different from the AdaBoost, GBRT fits the negative gradient of the loss function (L) obtained from the cumulative model of the previous iteration using the generated weak learners. Then, the negative gradient direction will be decreased by adding the obtained loss function to the weak learner. The process can be expressed as follows[45]:

$$F_m(x) = F_{m-1}(x) + \text{argmin}\sum_{i=1}^{n} \mathrm{L}[y_i, F_{m-1}(x_i) + h(x_i)] \tag{4}$$

$$\mathrm{g}_m(x) = -\frac{\partial L[y, F_{m-1}(x)]}{\partial F_{m-1}(x)} \tag{5}$$

where $h(x)$ is a basic learning function, and x is a vector of input features. g_m is the negative gradient of the loss function. The loss will be minimized when the m-th weak learner fits g_m of the loss function of the cumulative model[25]. The final gradient boosting regression tree is generated in the form of an ensemble of weak prediction models.

LightGBM is a framework for efficient implementation of the gradient boosting decision tee (GBDT) algorithm, which supports efficient parallel training with fast training speed and superior accuracy. Instead of segmenting the internal nodes of each tree using information gain as in traditional GBDT, LightGBM uses a gradient-based one-sided sampling (GOSS) method. In addition, LightGBM employs exclusive feature binding (EFB) to accelerate training without sacrificing accuracy[47].

Feature engineering

Feature engineering (FE) is the process of transforming raw data into features that better express the nature of the problem, enabling to improve the accuracy of model predictions on the invisible data. Data pre-processing, feature transformation, and feature selection are the main aspects of FE. Feature selection is the most important part of FE, which is to select useful features from a large number of features. It means that those features that are not relevant to the problem or are redundant with others need to be removed, and only the important features are retained in the end. Feature selection contains various methods such as correlation coefficient, principal component analysis, and mutual information methods. In this study, this process is done by the gray relation analysis (GRA) and Spearman correlation coefficient analysis, and the importance of features is calculated by the tree model. The implementation of data pre-processing and feature transformation will be described in detail in Section 3.1 and Section 3.2.

The basic idea of GRA is to determine the closeness of the connection according to the similarity of the geometric shapes of the sequence curves. The closer the shape of the curves, the higher the correlation of the corresponding sequences[23,48]. The method is used to analyze the degree of the influence of each factor on the results. The core is to establish a reference sequence according to certain rules, and then take each assessment object as a factor sequence and finally obtain their correlation with the reference sequence. In order to establish uniform evaluation criteria, variables need to be normalized according to Eq. 6 first due to the different attributes and units.

$$X_i(k) = \frac{x_i(k)}{\overline{x_i}} \tag{6}$$

Where, $X_i(k)$ represents the i-th value of factor k. The gray correlation between the reference series $X_0 = x_0(k)$ and the factor series $X_i = x_i(k)$ is defined as:

$$r(X_0, X_i) = \frac{1}{n}\sum_{k=1}^{n} r(X_0(k), X_i(k)) \tag{7}$$

$$r(X_0(k), X_i(k)) = \frac{\min\limits_i \min\limits_k |x_0(k) - x_i(k)| + \rho \min\limits_i \min\limits_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \min\limits_i \min\limits_k |x_0(k) - x_i(k)|} \tag{8}$$

Where, ρ is the discriminant coefficient and $\rho \in [0, 1]$, which serves to increase the significance of the difference between the correlation coefficients. Usually ρ is taken as 0.5.

The Spearman correlation coefficient is a parameter-free (distribution independent) test for measuring the strength of the association between variables. The Spearman correlation coefficient is solved according to the ranking of the original data[34]. Regardless of how the data of the two variables change and what distribution they fit, the order of the values is the only thing that is of interest. Two variables are significantly correlated if their corresponding values are ranked in the same or similar order within the group. In particular, if one variable is a strictly monotonic function of another variable, the Spearman Correlation Coefficient is equal to $+1$ or -1. The Spearman correlation coefficients of the variables R and S follow the equation:

$$P = \frac{\sum_{i=1}^{N}(R_i - \tilde{R})(S_i - \tilde{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \tilde{R})^2 \sum_{i=1}^{N}(S_i - \tilde{S})^2}} = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)} \tag{9}$$

Where, R_i and S_i are are the values of the variable R and S with rank i. $\tilde{R}$ and $\tilde{S}$ are the means of variables R and S, respectively. N is the total number of observations, and $d_i = R_i\text{-}S_i$, denoting the difference of variables in the same rank.

Explaining machine learning

The SHAP interpretation method is extended from the concept of Shapley value in game theory and aims to fairly distribute the players' contributions when they achieve a certain outcome

jointly[26]. SHAP values can be used in ML to quantify the contribution of each feature in the model that jointly provide predictions. The Shapley values of feature i in the model is:

$$Shap(i) = \sum_{k \subseteq M\{i\}} \frac{k!(N-k-1)!}{k!}[f_x(k \cup \{i\} - f_x(k)] \tag{10}$$

$$f_x(k) = E[f(x)|x_k] \tag{11}$$

Where, N denotes a subset of the features (inputs). $M\{i\}$ is the set of all possible combinations of features other than i. $E[f(x)|x_k]$ represents the expected value of the function on subset k. The prediction result y of the model is given in the following equation.

$$y = Shap_0 + \sum_{i=1}^{N} Shap(i) \tag{12}$$

Here, $shap_0$ is the average prediction of all observations and the sum of all SHAP values is equal to the actual prediction. Further, the absolute SHAP value reflects the strength of the impact of the feature on the model prediction, and thus the SHAP value can be used as the feature importance score[49,50].

The ALE plot describes the average effect of the feature variables on the predicted target. The most important property of ALE is that it is free from the constraint of variable independence assumption, which makes it gain wider application in practical environment. The key to ALE is to reduce a complex prediction function to a simple one that depends on only a few factors[29]. Then, the ALE plot is able to display the predicted changes and accumulate them on the grid. To quantify the local effects, features are divided into many intervals and non-central effects, which are estimated by the following equation.

$$\hat{\hat{f}}_{j,ALE(x)} = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^i \in N_j(k)} [f(Z_{kj}, x_j^i) - f(Z_{k-1}, x_j^i)] \tag{13}$$

Where, $Z_{i,j}$ denotes the boundary value of feature j in the k-th interval. $n_j(k)$ represents the sample size in the k-th interval. $i{:}x_j^i$ is the k-th sample point in the k-th interval, and x denotes the feature other than feature j. To make the average effect zero, the effect is centered as:

$$\hat{\hat{f}}_{j,ALE(x)} = \hat{\hat{f}}_{j,ALE(x)} - \frac{1}{n}\hat{\hat{f}}_{j,ALE}(x_j^i) \tag{14}$$

It means that the average effect is subtracted for each effect. Compared to the average predicted value of the data, the centered value could be interpreted as the main effect of the j-th feature at a certain point.

Performance metrics

In order to quantify the performance of the model well, five commonly used metrics are used in this study, including MAE, R^2, MSE, RMSE, and MAPE. Their equations are as follows.

$$MAE = \frac{\sum_{i=1}^{n} |P_i - T_i|}{n} \tag{15}$$

$$R^2 = 1 - \sum_{i=1}^{n} (P_i - T_i)^2 / \sum_{i=1}^{n} (P_i - \overline{T})^2 \tag{16}$$

$$MSE = \frac{\sum_{i=1}^{n} (P_i - T_i)^2}{n} \tag{17}$$

$$RMSE = \sqrt{MSE} \tag{18}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{|P_i - T_i|}{T_i} \right| \tag{19}$$

Where, T_i represents the actual maximum pitting depth, the predicted value is P_i, and n denotes the number of samples. R^2 reflects the linear relationship between the predicted and actual value and is better when close to 1. MSE, RMSE, MAE, and MAPE measure the relative error between the predicted and actual value. The values of the above metrics are desired to be low.

Implementation methodology

Step 1: Pre-processing. Pre-processing of the data is an important step in the construction of ML models. Although some of the outliers were flagged in the original dataset, more precise screening of the outliers was required to ensure the accuracy and robustness of the model. CV and box plots of data distribution were used to determine and identify outliers in the original database. In addition, the type of soil and coating in the original database are categorical variables in textual form, which need to be transformed into quantitative variables by one-hot encoding in order to perform regression tasks.

Step 2: Model construction and comparison. After pre-processing, 200 samples of the data were chosen randomly as the training set and the remaining 40 samples as the test set. The RF, AdaBoost, GBRT, and LightGBM methods introduced in the previous section and ANN models were applied to the training set to establish models for predicting the dmax of oil and gas pipelines with default hyperparameters. Then a promising model was selected by comparing the prediction results and performance metrics of different models on the test set.

Step 3: Optimization of the best model. The best model was determined based on the evaluation of **step 2**. However, the performance of an ML model is influenced by a number of factors. Apart from the influence of data quality, the hyperparameters of the model are the most important. In this step, the impact of variations in the hyperparameters on the model was evaluated individually, and the multiple combinations of parameters were systematically traversed using grid search and cross-validated to determine the optimum parameters. This optimized best model was also used on the test set, and the predictions obtained will be analyzed more carefully in the next step.

Step 4: Model visualization and interpretation. In the most of the previous studies, different from traditional mathematical formal models, the optimized and trained ML model does not have a simple expression. The model is saved in the computer in an extremely complex form and has poor readability. In this study, this complex tree model was clearly presented using visualization tools for review and application. After completing the above, the SHAP and ALE values of the features were calculated to provide a global and localized interpretation of the model, including the degree of contribution of each feature to the prediction, the influence pattern, and the interaction effect between the features.

DATA AVAILABILITY

The original dataset for this study is obtained from Prof. F. Caleyo's dataset (https://doi.org/10.5006/1.3318290). More calculated data and python code in the paper is available via the corresponding author's email.

CODE AVAILABILITY

The machine learning approach framework used in this paper relies on the python package. We are happy to share the complete codes to all researchers through the corresponding author.

Received: 19 September 2022; Accepted: 7 January 2023;
Published online: 02 February 2023

REFERENCES

1. Wasim, M. & Djukic, M. B. External corrosion of oil and gas pipelines: A review of failure mechanisms and predictive preventions. *J. Nat. Gas. Sci. Eng.* **100**, 104467 (2022).
2. El Amine Ben Seghier, M. et al. Prediction of maximum pitting corrosion depth in oil and gas pipelines. *Eng. Fail Anal.* **112**, 104505 (2020).
3. Askari, M., Aliofkhazraei, M. & Afroukhteh, S. A comprehensive review on internal corrosion and cracking of oil and gas pipelines. *J. Nat. Gas. Sci. Eng.* **71**, 102971 (2019).
4. Shuai, Y., Wang, X. & Cheng, Y. F. Buckling resistance of an X80 steel pipeline at corrosion defect under bending moment. *J. Nat. Gas. Sci. Eng.* **93**, 104016 (2021).
5. Shuai, Y., Wang, X. & Cheng, Y. F. Modeling of local buckling of corroded X80 gas pipeline under axial compression loading. *J. Nat. Gas. Sci. Eng.* **81**, 103472 (2020).
6. Xu, M. et al. Effect of pressure on corrosion behavior of X60, X65, X70, and X80 carbon steels in water-unsaturated supercritical CO_2 environments. *Int. J. Greenh. Gas Control* **51**, 357–368 (2016).
7. Singh, M., Markeset, T. & Kumar, U. Some philosophical issues in modeling corrosion of oil and gas pipelines. *Int. J. Syst. Assur.* **5**, 55–74 (2014).
8. Lam, C. & Zhou, W. Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database. *Int. J. Pres. Vessel. Pip.* **145**, 29–40 (2016).
9. Rippon, I. J. A. Corrosion management for an offshore sour gas pipeline system. (NACE International, Houston, Texas, 2005).
10. PENG, C. et al. Corrosion and pitting behavior of pure aluminum 1060 exposed to Nansha Islands tropical marine atmosphere. *T. Nonferr. Metal. Soc.* **32**, 448–460 (2022).
11. Wei, W. et al. In-situ characterization of initial marine corrosion induced by rare-earth elements modified inclusions in Zr-Ti deoxidized low-alloy steels. *J. Mater. Res. Technol.* **9**, 1412–1424 (2020).
12. Tran, N., Nguyen, T., Phan, V. & Nguyen, D. A machine learning-based model for predicting atmospheric corrosion rate of carbon steel. *Adv. Mater. Sci. Eng.* **2021**, 1–25 (2021).
13. Song, X. et al. Multi-factor mining and corrosion rate prediction model construction of carbon steel under dynamic atmospheric corrosion environment. *Eng. Fail Anal.* **134**, 105987 (2022).
14. Ren, C., Qiao, W. & Tian, X. Natural gas pipeline corrosion rate prediction model based on BP neural network. *Fuzzy Inf. Eng. Oper. Res.* **147**, 449–455 (2012).
15. Xie, M., Li, Z., Zhao, J. & Pei, X. A prognostics method based on back propagation neural network for corroded pipelines. *Micromachines* **12**, 1568 (2021).
16. Liao, K., Yao, Q., Wu, X. & Jia, W. A numerical corrosion rate prediction method for direct assessment of wet gas gathering pipelines internal corrosion. *Energies* **5**, 3892–3907 (2012).
17. Li, X., Jia, R., Zhang, R., Yang, S. & Chen, G. A KPCA-BRANN based data-driven approach to model corrosion degradation of subsea oil pipelines. *Reliab. Eng. Syst. Saf.* **219**, 108231 (2022).
18. Ossai, C. I. Corrosion defect modelling of aged pipelines with a feed-forward multi-layer neural network for leak and burst failure estimation. *Eng. Fail Anal.* **110**, 104397 (2020).
19. Abbas, M. H., Norman, R. & Charles, A. Neural network modelling of high pressure CO_2 corrosion in pipeline steels. *Process. Saf. Environ.* **119**, 36–45 (2018).
20. Hernández, S., Nešić, S. & Weckman, G. R. Use of Artificial Neural Networks for predicting crude oil effect on CO_2 corrosion of carbon steels. *Corrosion* **62**, 467–482 (2005).
21. Ossai, C. I. & Data-Driven, A. Machine learning approach for corrosion risk assessment—a comparative study. *Big Data Cogn. Comput.* **3**, 28 (2019).
22. De Masi, G. et al. Machine learning approach to corrosion assessment in subsea pipelines. (OCEANS 2015 - Genova, Genova, Italy, 2015).
23. Luo, Z., Hu, X., & Gao, Y. Corrosion research of wet natural gathering and transportation pipeline based on SVM. (ICPTT 2013).
24. Zhang, W. D., Shen, B., Ai, Y. B. & Yang, B. Gas pipeline corrosion prediction based on modified support vector machine and unequal interval model. *Appl. Mech. Mater.* **373-375**, 1987–1994 (2013).
25. Ben Seghier, M. E. A., Höche, D. & Zheludkevich, M. Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques. *J. Nat. Gas Sci. Eng.* **99**, 104425 (2022).
26. Ekanayake, I. U. & Meddage, D. P. P. U. Rathnayake. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud. Constr. Mater.* **16**, e1059 (2022).
27. Xu, F. et al. *Natural Language Processing and Chinese Computing 563-574*. Lecture Notes in Computer Science, Vol. 11839 (Springer, 2019).
28. Wang, Z., Zhou, T. & Sundmacher, K. Interpretable machine learning for accelerating the discovery of metal-organic frameworks for ethane/ethylene separation. *Chem. Eng. J.* **444**, 136651 (2022).
29. Liu, K. et al. Interpretable machine learning for battery capacities prediction and coating parameters analysis. *Control. Eng. Pract.* **124**, 105202 (2022).
30. Velázquez, J., Caleyo, F., Valor, A, & Hallen, J. M. Technical note: field study—pitting corrosion of underground pipelines related to local soil and pipe characteristics. *Corrosion*. **66**, 016001-1–016001-5 (2010).
31. Kim, C., Chen, L., Wang, H. & Castaneda, H. Global and local parameters for characterizing and modeling external corrosion in underground coated steel pipelines: a review of critical factors. *J. Pipeline Syst. Eng.* **1**, 17–35 (2021).
32. Dai, M., Liu, J., Huang, F., Zhang, Y. & Cheng, Y. F. Effect of cathodic protection potential fluctuations on pitting corrosion of X100 pipeline steel in acidic soil environment. *Corros. Sci.* **143**, 428–437 (2018).
33. Zhang, B. et al. Unmasking chloride attack on the passive film of metals. *Nat. Commun.* **9**, 2559 (2018).
34. Zhi, Y. et al. Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model. *Corros. Sci.* **178**, 109084 (2021).
35. Song, Y. et al. Effects of chloride ions on corrosion of ductile iron and carbon steel in soil environments. *Sci. Rep.* **7**, 6865 (2017).
36. Wang, Y. et al. Effect of pH and chloride on the micro-mechanism of pitting corrosion for high strength pipeline steel in aerated NaCl solutions. *Appl. Surf. Sci.* **349**, 746–756 (2015).
37. Wasim, M., Shoaib, S., Mujawar, M., Inamuddin & Asiri, A. M. Factors influencing corrosion of metal pipes in soils. *Environ. Chem. Lett.* **16**, 1–19 (2018).
38. Sani, F. M. The effect of bacteria and soil moisture content on external corrosion of buried pipelines. (NACE International, Virtual, 2021).
39. Chen, J. et al. Impact of soil composition and electrochemistry on corrosion of rock-cut slope nets along railway lines in China. *Sci. Rep.* **5**, 14939 (2015).
40. Bash, L. A. R. Pipe-to-soil potential measurements, the basic science. (NACE International, New Orleans, Louisiana, 2008).
41. Li, X. & Castaneda, H. Damage evolution of coated steel pipe under cathodic-protection in soil. *Anti-Corros. Methods Mater.* **64**, 118–126 (2017).
42. Amaya-Gómez, R., Bastidas-Arteaga, E., Muñoz, F. & Sánchez-Silva, M. Statistical soil characterization of an underground corroded pipeline using in-line inspections. *Metals* **11**, 292 (2021).
43. Apley, D., Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc.* **82**, 1059–1086 (2020).
44. Salami, B. A., Rahman, S. M., Oyehan, T. A., Maslehuddin, M. & Al Dulaijan, S. U. Ensemble machine learning model for corrosion initiation time estimation of embedded steel reinforced self-compacting concrete. *Measurement* **165**, 108141 (2020).
45. Feng, D., Wang, W., Mangalathu, S., Hu, G. & Wu, T. Implementing ensemble learning methods to predict the shear strength of RC deep beams with/without web reinforcements. *Eng. Struct.* **235**, 111979 (2021).
46. Cao, Y., Miao, Q., Liu, J. & Gao, L. Advance and prospects of AdaBoost algorithm. *Acta Autom. Sin.* **39**, 745–758 (2013).
47. Wen, X., Xie, Y., Wu, L. & Jiang, L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accid. Anal. Prev.* **159**, 106261 (2021).
48. Liu, S., Cai, H., Cao, Y. & Yang, Y. Advance in grey incidence analysis modelling. (IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, USA, 2011).
49. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Environ. Urban. Syst.* **96**, 101845 (2022).
50. Mamun, O., Wenzlick, M., Sathanur, A., Hawk, J. & Devanathan, R. Machine learning augmented predictive and generative model for rupture life in ferritic and austenitic steels. *Npj Mater. Degrad.* **5**, 1–10 (2021).

ACKNOWLEDGEMENTS

This research was financially supported by the National Natural Science Foundation of China (No. 52174007, No. 51801167, No. 52001264), the Opening Project of Material Corrosion and Protection Key Laboratory of Sichuan province (No. 2022CL04), and Project of Sichuan Department of Science and Technology (No. 23SYSX0127). The authors thank Prof. F. Caleyo and his team for making the complete database publicly available.

AUTHOR CONTRIBUTIONS

Y.S.: methodology, investigation, data duration, software, validation, visualization, and writing – original draft. Q.W.: conceptualization, methodology, investigation, resources, data curation, writing – review & editing, supervision, and funding

acquisition. X.Z.: methodology and data curation. L.D.: methodology, data curation, and funding acquisition. S.B.: methodology and resources. D.Z.: methodology, funding acquisition, and supervision. Z.Z.: conceptualization, methodology, and funding acquisition. H.Z.: conceptualization, review, and resources. Y.X.: methodology, funding acquisition, and supervision.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-023-00324-x.

Correspondence and requests for materials should be addressed to Qinying Wang, Lijin Dong or Yuchen Xi.

Reprints and permission information is available at http://www.nature.com/reprints

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