

ARTICLE OPEN



Data-driven pitting evolution prediction for corrosion-resistant alloys by time-series analysis

Xue Jiang¹, Yu Yan²✉ and Yanjing Su²

Corrosion initiation and propagation are a time-series problem, evolving continuously with corrosion time, and future pitting behavior depends closely on the past. Predicting localized corrosion for corrosion-resistant alloys remains a great challenge, as macroscopic experiments and microscopic theoretical simulations cannot couple internal and external factors to describe the pitting evolution from a time dimension. In this work, a data-driven method based on time-series analysis was explored. Taking cobalt-based alloys and duplex stainless steels as the case scenario, a corrosion propagation model was built to predict the free corrosion potential (E_{corr}) using a long short-term memory neural network (LSTM) based on 150 days of immersion testing in saline solution. Compared to traditional machine learning methods, the time-series analysis method was more consistent with the evolution of ground truth in the E_{corr} prediction of the subsequent 70 days' immersion, illustrating that time-series dependency of pitting propagation could be captured and utilized.

npj Materials Degradation (2022)6:92; <https://doi.org/10.1038/s41529-022-00307-4>

INTRODUCTION

Corrosion initiation and propagation have the characteristics of randomness, dynamics, and the coupling of internal alloy factors and external environmental factors^{1–4}. This involves the rupture or dissolution of the passive film in corrosive environments, the dissolution rate of the metal on the pitting surface and the dissolved metal at different moments, as well as the diffusion rate of cations in the corrosive solution^{5,6}. These factors jointly influence the local chemical environment and the dynamic evolution of pit morphology with the corrosion time, which in turn determine the design of the material composition and structure. Researchers have carried out numerous experimental and theoretical studies on pitting mechanisms, describing pitting evolution as either discrete and macroscopic or continuous and microscopic^{7–9}. From the perspective of alloy service time, pitting initiation and propagation are closely related to the interactions between local corrosion conditions and material structure, which can evolve and accumulate continuously with corrosion time. To date, making accurate predictions of pitting propagation for corrosion-resistant alloys requires capturing past corrosion phenomena to estimate future possibilities, which has remained a great challenge¹.

Artificial intelligence (AI) and machine learning (ML) can capture complex relationships between multi-dimension factors and targets; thus, promoting new materials and insight discovery^{10–15}. Since the 1980s, researchers have applied artificial neural networks, random forests, and other machine learning algorithms to predict the uniform corrosion of materials and design new corrosion-resistant materials, making notable progress in solving the multi-factor coupling corrosion problem^{16–21}. This has helped to predict the corrosion rate of low alloy steel and carbon steel, analyze the important factors that affect the corrosion rate, and forecast the local corrosion behavior of Co-based alloys under different compositions, preparation processes, temperatures, static corrosion environments, and corrosion times^{22–26}. Coelho et al.

provides a data-oriented overview of the rapidly growing research field covering ML applied to predicting electrochemical corrosion, which highlights assessing the predictive power of different approaches and elaborate on the current status of regression modeling for various corrosion topics²⁷. Sharma et al. have employed Random Forest method to model measurements of corrosion rates of carbon steel as a function of time when corrosion inhibitors are added in different dosage and dose-schedules²⁸. However, traditional statistical analysis methods, such as support vector machine, random forest and gradient boosting regression default to the assumption of independent and identical distribution among the data samples. For the dynamic process of material service, traditional analysis methods cannot use the inherent time-series relationship between the data samples. Time-series data analysis methods, such as the well-known long short-term memory (LSTM) neural network, can perform dependency mining on sequence data, and learn functions that map a sequence of past observations as an input to output observation²⁹. By introducing input gates, forgetting gates, output gates, and memory units, the information in the memory units can be maintained, updated, or forgotten at different moments to solve the problems of gradient disappearance and sequence dependencies under long-term time series³⁰. Thus, the time-series data analysis method can provide a prospect for the evolution prediction of pitting propagation.

In this work, using cobalt-based alloys and duplex stainless steels as the case scenario, we explored a machine learning method based on the time-series analysis of an LSTM neural network. We constructed a pitting propagation model of free corrosion potential (E_{corr}) based on a 150-day immersion test in saline. Traditional machine learning algorithms were compared with the LSTM, which were further applied to estimate the E_{corr} of the following 70 days in future immersion. The LSTM model was more consistent with the evolution trend of the ground truth outlined by experimentation, illustrating that the time-series

¹Beijing Advanced Innovation Center for Materials Genome Engineering, Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, 100083 Beijing, China. ²Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, 100083 Beijing, China. ✉email: yanyu@ustb.edu.cn

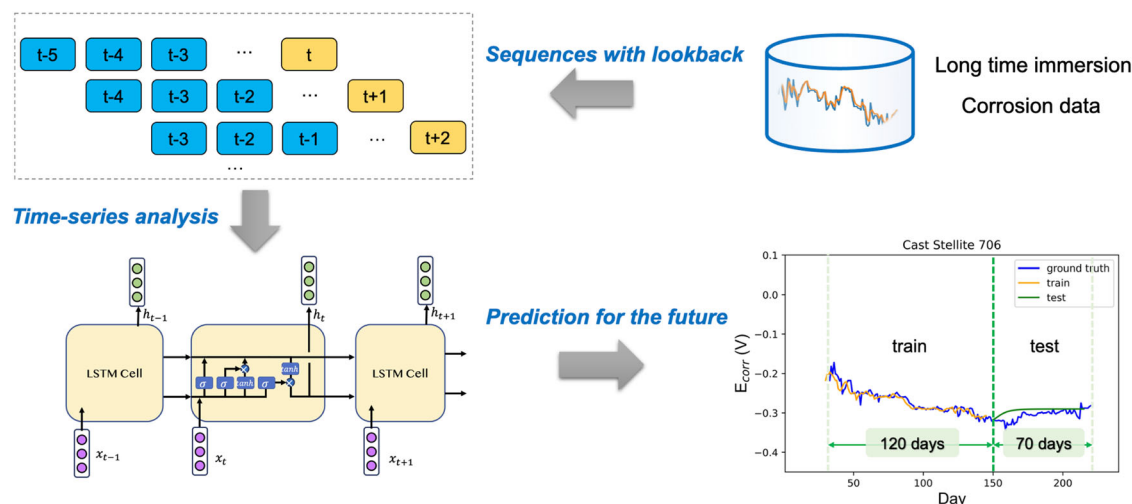


Fig. 1 The workflow diagram of machine learning. The collected dataset are transformed into sequences with appropriate lookback to maintain the information of former moment. Then LSTM neural network is trained to capture the dependences, and it is finally used to conduct future prediction for the following days' pitting.

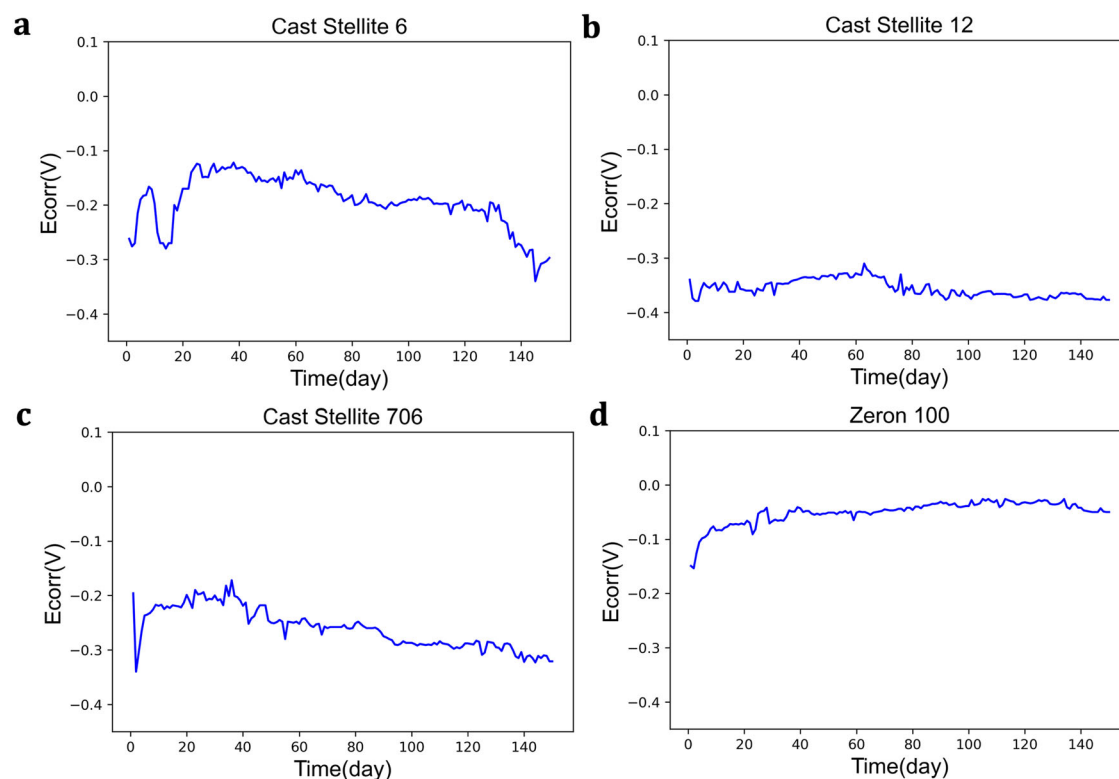


Fig. 2 The visualization of original dataset for different alloys. The free corrosion potential (E_{corr}) values measured during the immersion period of 150 days: **a** Cast Stellite 6, **b** Cast Stellite 12, **c** Cast Stellite 706, and **d** Zeron 100.

analysis method could capture the sequence dependencies of pitting propagation under a long-term time series attributed to the network structure. This may pave a promising approach for other material services and lifetime behavior predictions.

RESULTS AND DISCUSSION

Machine learning strategy

The machine learning strategy starts with data collection, goes through sequence transformation, time-series analysis, and finally

predicts future pitting behavior (Fig. 1). The whole dataset was prepared by long-time immersion for four different alloys of Stellite 6, Stellite 12, Stellite 706, and Zeron 100. Long-term immersion tests were conducted in 3.5 wt. % NaCl solution at 18 °C for 150 days. E_{corr} was measured every day during the immersion period for half a year.

Figure 2 shows the E_{corr} values of the samples versus the immersion time for 150 days. During immersion, visible pitting was first observed around 30 days for all alloy samples. Thus, we manually removed the first 30-day data, because pitting was in the initiation stage during this time, and E_{corr} fluctuated greatly.

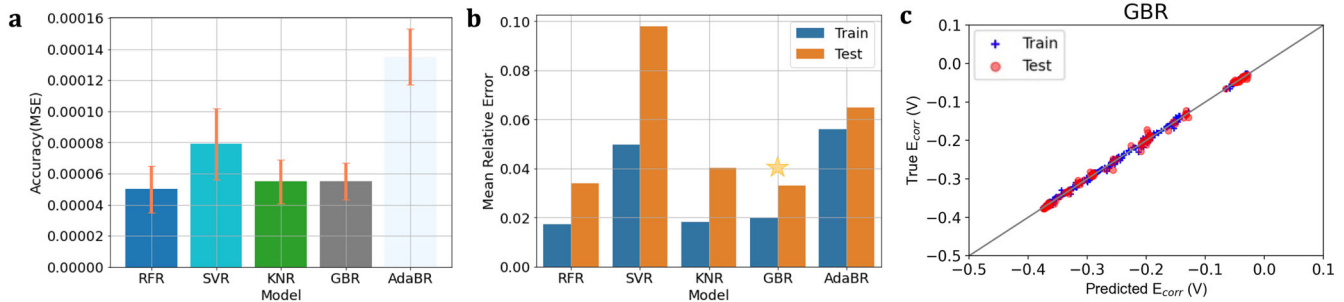


Fig. 3 Evaluation for traditional machine learning models. Parameter tuning and model selection of the traditional machine learning models: **a** The mean squared errors with variances (error bars) of the parameter tuning process for RFR, SVR, KNR, GBR, and AdaBR; **b** The mean relative errors of machine learning models on the training set and testing set separately; **c** Diagonal scatter plot of the ground truth versus the predicted E_{corr} by GBR during training and testing, respectively. E_{corr} free corrosion potential, SVR support vector regression, KNR k-nearest neighbor regression, GBR gradient boosting regression, RFR random forest regression, AdaBR AdaBoost regression.

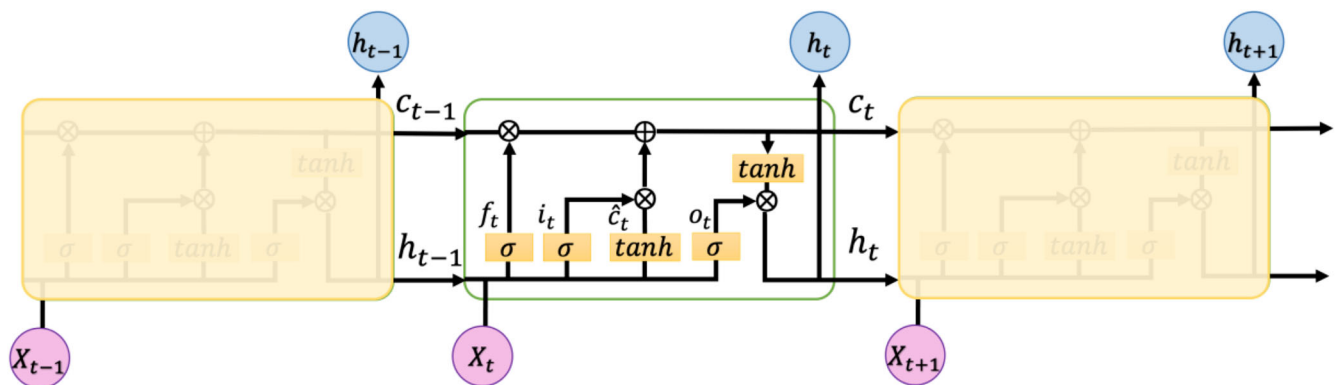


Fig. 4 The internal structure of LSTM, including the forget gate, input gate, and output gate. The pink circle represents the input of different moment, the blue circle represents the hidden state of different moment.

Therefore, the dataset consisted of 480 entries in total, including the composition of each alloy, Vickers hardness (HV), and standard deviation of HV, as well as the immersion time of 120 days (days from 30 to 150), along with the target property of E_{corr} . We shared the dataset to the Materials Genome Engineering Database to facilitate the data for further reuse, and the dataset can be found by this link (<https://www.mgedata.cn/search/#/153870/1064>).

Traditional machine learning models

Firstly, traditional machine learning algorithms have been used, including support vector regression (SVR), k-nearest neighbor regression (KNR), gradient boosting regression (GBR), random forest regression (RFR), and AdaBoost regression (AdaBR)³¹. Machine learning algorithms regarded immersion time as one of the conditionally independent feature variables, capturing the relationship between the different chemical compositions, immersion times, and E_{corr} . The chemical composition, HV, standard deviation of HV, and immersion time were utilized as the inputs, with the target property E_{corr} as the outputs. Before model training, parameter tuning was performed to obtain the optimal model parameters. The whole dataset was randomly split with 80% as the training set (384 data entries) and 20% left as the testing set (96 data entries). Parameter tuning was employed on the training set using grid search by five-fold cross-validation with randomly selected hyperparameters for each machine learning algorithm.

We used mean squared error (MSE) as a metric for the error between the ground truth and predicted value under each hyperparameter. For five-fold cross-validation, the mean and variance of the five-fold MSEs were calculated to evaluate the

comprehensive performance of all hyperparameters for each algorithm. Figure 3a shows the mean and variance values of the optimal hyperparameter for each algorithm. RFR performed the lowest mean MSE, while KNR and GBR also had a relatively low MSE; however, GBR exhibited a smaller variance, indicating that the GBR model would be more stable. Moreover, to evaluate the generalization capability intuitively, the models were also tested on the hold-out testing set with the optimal hyperparameter for each algorithm by the mean relative error (MRE) on both the training and testing sets (Fig. 3b).

Although RFR exhibited better MRE on the training set than the other models, the difference in MRE between the training and unseen testing set was larger than GBR. Thus, GBR with a test MRE of 3.3% proved to have a better generalization ability and could be used for constructing the E_{corr} prediction model of pitting propagation. Figure 3c gives the diagonal scatter plot of the ground truth versus the predicted E_{corr} by GBR during training and testing respectively, illustrating good fitting.

LSTM neural network

To maintain the information of former moment, LSTM neural network was further trained to capture the time dependences. The LSTM neural network consists of a gated recurrent neural network (RNN) that can account for long-time dependencies, can perform dependency mining on sequence data, and learn a function that maps the sequence of past observations as an input to output observation³⁰. LSTM can use memory cell c_t to remove or add information. c_t is composed of the input gate i_t , forget gate f_t , and output gate o_t (Fig. 4).

The LSTM model was implemented by TensorFlow and Keras after transforming the original dataset to a set of training sequences with a sliding window, namely the lookback window^{32,33}. For E_{corr} prediction of long-time immersion, lookback determined how many days of immersion data would be used to predict the E_{corr} of the next day. As shown in Fig. 5a, lookback = 5 meant using the previous 5 days' E_{corr} to predict the E_{corr} on day 6, and the window kept moving to the right. The larger the lookback window, the longer the model would capture the sequence dependencies; however, the fewer training sequences that would be generated based on 120 days (days 30–150) of data, and the model would be prone to overfitting. Thus, the smaller the window, the more training sequences that could be constructed; however, long-term sequence dependencies could not be captured, which required a tradeoff. We carried out parameter tuning of the lookback window for the LSTM neural network, as shown in Fig. 5b. For each lookback length, the LSTM neural network was trained on the first 80% sequences as the training set and tested on the remaining 20% sequences as the testing set to evaluate the generalization ability. Figure 5b shows the MSEs of the different lookbacks for cast Stellite 6 on the training and testing set. We concluded that lookback = 5 was optimal, as the others resulted in overfitting.

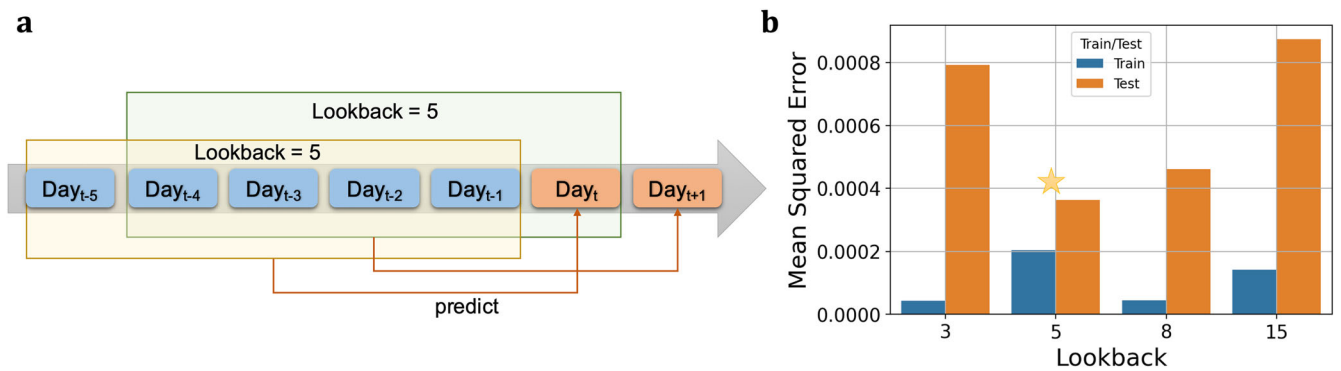


Fig. 5 Schematic workflow and evaluation for LSTM neural network. The schematic workflow and parameter tuning for the LSTM neural network: **a** Schematic workflow used by the LSTM model with the time-series immersion dataset; **b** Taking Cast Stellite 6 as an example, the mean squared errors of the different lookbacks for the LSTM model on the training and testing set are shown. LSTM long short-term memory.

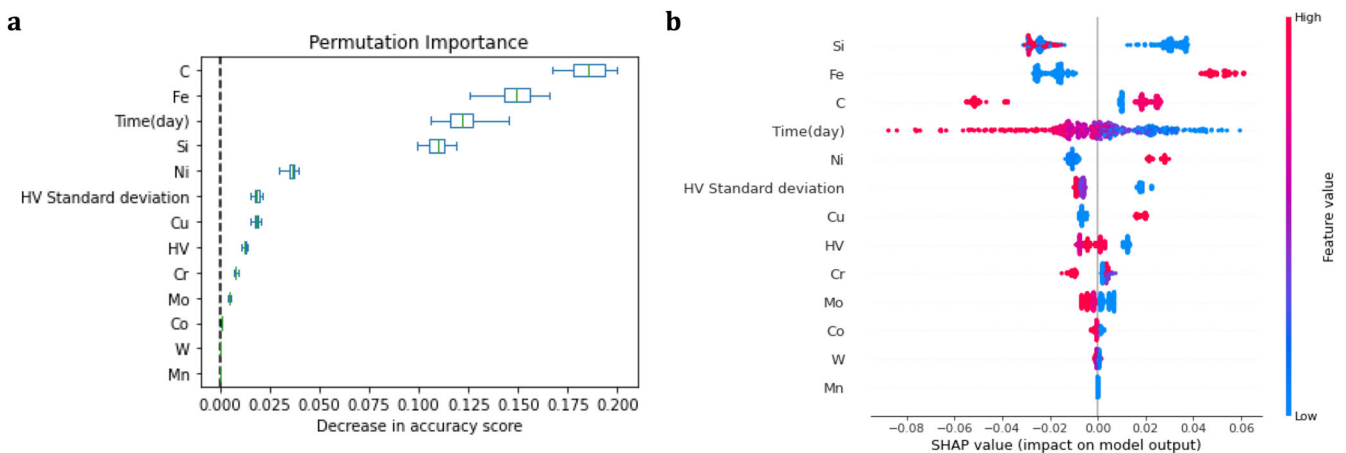


Fig. 6 Feature importance analysis for GBR model. Feature importance as given by the permutation importance and SHAP value based on the GBR E_{corr} model: **a** Box plot of the feature contributions by permutation importance with the minimum and maximum score, lower and upper quartile, and mean value of various features versus its decrease in accuracy score of models; **b** Distribution plot of the feature contributions by the SHAP value. High feature value is colored with red and low feature value is colored with blue. GBR gradient boosting regression, SHAP Shapley Additive exPlanations.

Feature importance

To explore the contribution and importance of the alloy elements and immersion time in predicting E_{corr} , feature importance analysis was carried out by two different methods, namely the permutation feature importance and Shapley Additive exPlanations (SHAP) values^{34,35}. The permutation feature importance was used to reveal the effect of the feature to the target by shuffling a single feature. Therefore, a decrease in the model score indicates that the model is highly dependent on the feature, especially for nonlinear estimators. Figure 6a shows the box plot of the permutation importance distribution based on the GBR E_{corr} model by repeating 50 times, where the features are shown in descending order according to their contribution to decreasing the accuracy score (MSE). The components of C, Fe, Si, and the immersion time acted as more important when they were shuffled, and they showed a great drop in the MSE of the E_{corr} model. Furthermore, as a game theoretic approach, SHAP can interpret the output of machine learning models by the classic Shapley value. A feature's positive SHAP value enhanced the properties, while conversely, a negative SHAP value for a feature weakened the properties.

Figure 6b shows how each feature affected the GBR model output of E_{corr} , where the features are shown in descending order according to their importance by the average absolute value of all

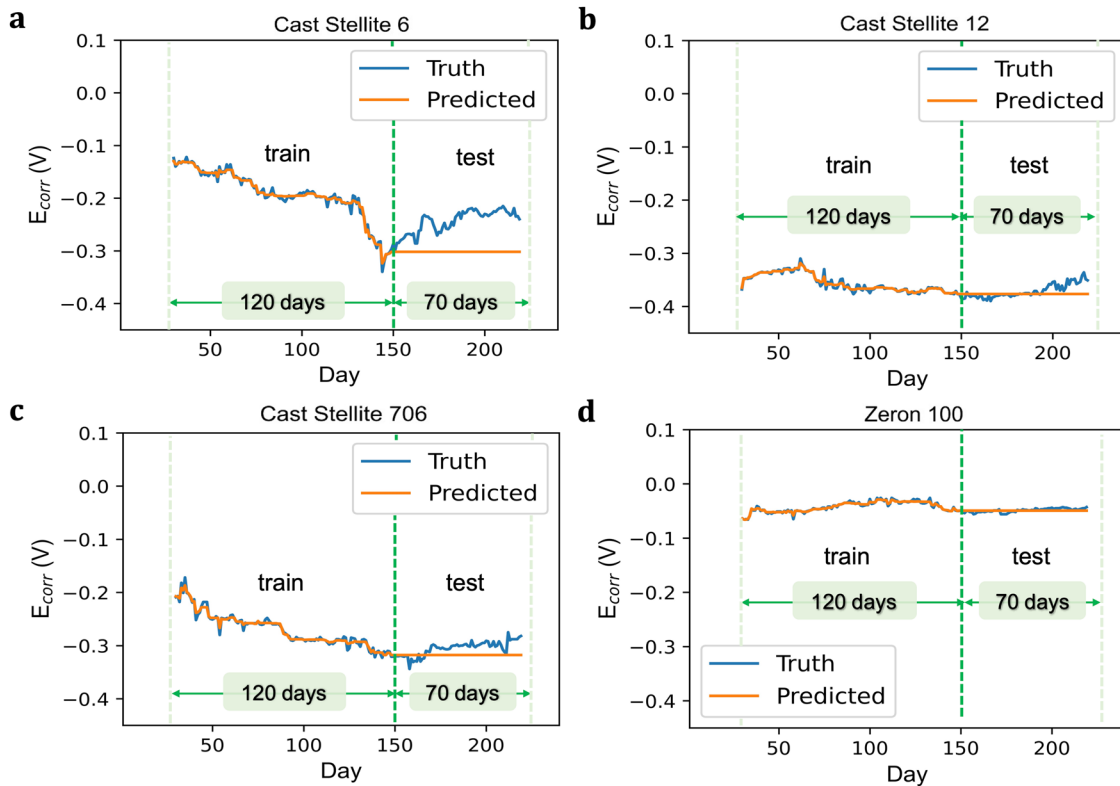


Fig. 7 The prediction of following 70 days' immersion by GBR model. Prediction of the following 70 days of immersion by the GBR model: **a** Cast Stellite 6, **b** Cast Stellite 12, **c** Cast Stellite 706, and **d** Zeron 100. GBR gradient boosting regression, E_{corr} free corrosion potential.

features' SHAP values. The components of Si, Fe, C, and immersion time also acted as the most important factors for E_{corr} prediction. For element Fe in Fig. 6b, the SHAP value was positive when increasing the component of Fe, which indicated that the addition of Fe would enhance E_{corr} (the output of the model was kept stable). For the immersion time, it was obvious that a longer immersion time would weaken E_{corr} (the output of the model tended to descend).

Overall, the two feature importance analysis methods based on different mechanisms both considered Fe, C, Si, and the immersion time to be most important for E_{corr} , although they differed slightly in the importance order of Si. For the duplex stainless steel Zeron 100, Fe was the main element and C was 0.03%, while for the three cobalt-based alloys, Co was the main element, the content of Fe element accounted for 2–3%, while the content of C was greater than 1%. The pitting corrosion resistance of Zeron 100 was significantly better than the cobalt-based alloys. Therefore, the composition of the alloys was significantly different, resulting in the higher importance of these three elements. It is worth noting that the immersion time had an obvious influence on the E_{corr} prediction by GBR (from Fig. 6b), indicating that the local corrosion states of the previous sequence were inherited into the prediction of the subsequent sequence and the cumulative effect over time dimension needs to be utilized.

The prediction following 70 days of immersion

To validate the analysis results and estimate the E_{corr} of unseen immersion, additional 70-day immersion tests were carried out based on the previous experiment and the E_{corr} was measured each day. We utilized the pre-built E_{corr} models of GBR and LSTM to predict the following 70 days of pitting evolution. Figure 7 describes the training (120 days) and testing (70 days) stages by the GBR model for cast Stellite 6, cast Stellite 12, cast Stellite 706,

and Zeron 100. Although the predicted values and the ground truths fit well in the training stage, in the subsequent 70-day testing stage, the GBR model gave fixed predicted values of the four alloys, which did not match the evolution trend of E_{corr} . This indicated that traditional statistical analysis methods such as GBR defaulted to the assumption of independent and identical distribution among the data samples. For dynamic pitting evolution, it could not make use of the inherent time-series relationship between the data samples.

Figure 8 shows the prediction of the following 70 days of immersion by the LSTM model. Compared to the GBR model, the time-series model fit the trend of E_{corr} well, especially in the testing stage (green plot in Fig. 8). The absolute deviation between the predicted value and the ground truth was somewhat large for cast Stellite 6, which was mainly due to the significant increase in E_{corr} around day 150 (the dividing line between the training and testing stages). For the cast Stellite 12, 706, and Zeron 100 alloys, the absolute deviations between the predicted and ground truths at the testing stage were smaller. This indicated that the LSTM model could capture the inherent time-series relationship between the data samples and could be utilized in long-term pitting evolution prediction.

In summary, by taking cobalt-based alloys and duplex stainless steel as the case scenario, we explored a time-series-based machine learning method for predicting the pitting evolution behavior by LSTM. We constructed a pitting evolution model to predict E_{corr} under immersion testing in saline during 150 days in the propagation stage. The traditional machine learning model GBR was compared with LSTM, which did not fit well in the evolution trend of E_{corr} during the subsequent 70-day unseen testing stage. The LSTM model was more consistent with the evolution trend of the ground truth from experimentation, which illustrated that this method could capture the inherent time-series

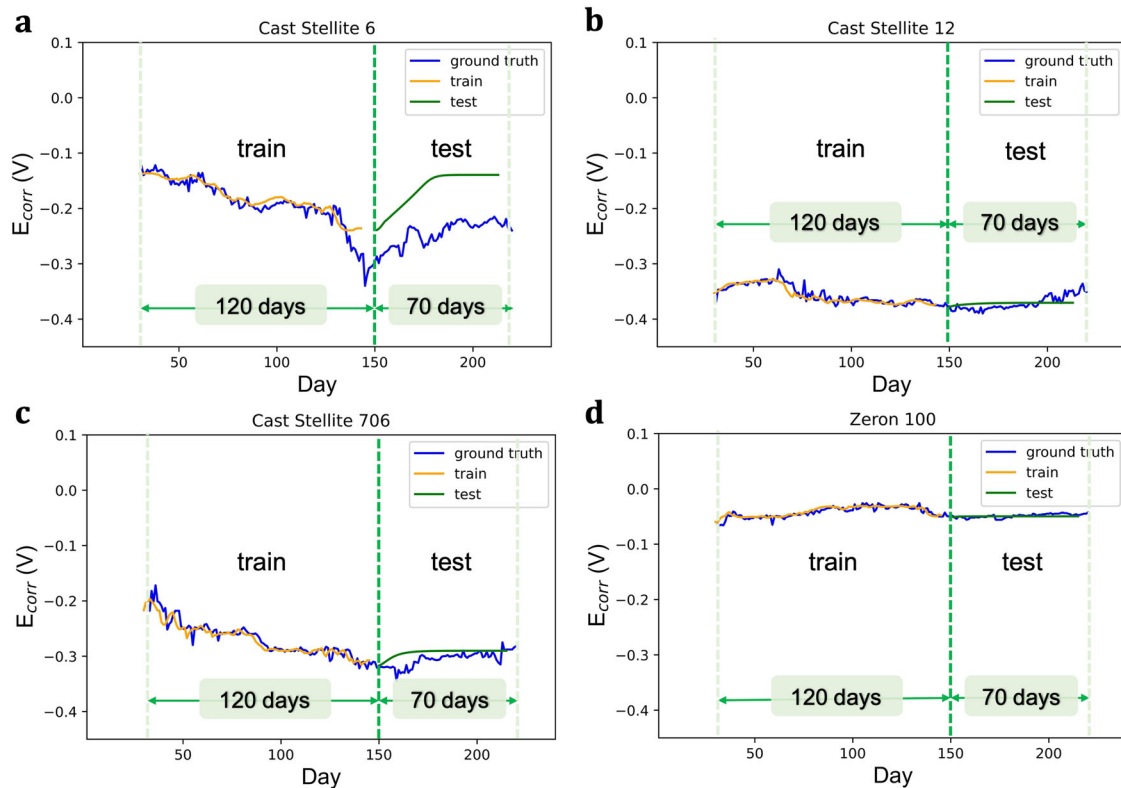


Fig. 8 The prediction of following 70 days' immersion by LSTM model. Prediction of the following 70 days of immersion by the LSTM model: **a** Cast Stellite 6, **b** Cast Stellite 12, **c** Cast Stellite 706, and **d** Zeron 100. LSTM long short-term memory, E_{corr} free corrosion potential.

Table 1. Chemical compositions (wt.%) of the experimental material.

Alloys	Co	Cr	W	Mo	Fe	C	Ni	Si	Mn	Cu	HV	HV standard deviation
Cast Stellite 6	58.4	28.9	4.5	0	2.4	1.2	2.5	1.1	1	0	496.6	25.145
Cast Stellite 12	52.6	29	8.5	0	3	1.4	3	1.5	1	0	504.2	40.25
Cast Stellite 706	50.3	27	0	5	2	1.2	3	10.5	1	0	465	21.722
Zeron 100	0	26	1	4	57.47	0.03	8.5	1	1	1	362.5	13.199

relationship between the data samples and could be utilized in long-term pitting evolution prediction. This work may provide a promising avenue for time-series sequence prediction, which could be further generalized for other material service and lifetime behavior predictions.

METHODS

Data collection

All of the data used in this work were obtained by experiments in laboratory, including four different alloys: Stellite 6, Stellite 12, Stellite 706, and Zeron 100. The nominal chemical compositions of each material are shown in Table 1.

The samples were prepared by casting and were embedded in epoxy resin with a wire soldered to the rear. The surfaces were polished using 80, 240, 600, and 1200 grit SiC paper and then polished using 6-micron diamond paste. The hardness of these materials was also measured up to 50 times for each sample. Long-term immersion tests were conducted in 3.5 wt.% NaCl

solution at 18 °C for 150 days. The images of the samples under different immersion stages were given in Supplementary Fig. 1.

Machine learning metrics

Mean squared error (MSE) was used as a metric for the error between the ground truth and predicted value under each hyperparameter, and MSE could be calculated by

$$MSE = \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N}, \quad (1)$$

where N is the total number of the dataset, and y_i and \hat{y}_i represent the ground truth and predicted value of the i th data entry, respectively.

Moreover, to evaluate the generalization capability intuitively, mean relative error (MRE) was also calculated on both the training and testing sets. MRE is defined by Eq. (2):

$$MRE = \frac{\sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|}{N}, \quad (2)$$

where N is the total number of the dataset, and y_i and \hat{y}_i represent the ground truth and predicted value of the i th data entry, respectively.

Traditional machine learning models

Support vector regression (SVR), k-nearest neighbor regression (KNN), gradient boosting regression (GBR), random forest regression (RFR), and AdaBoost regression (AdaBR) were used based on the Scikit-learn machine learning library. Parameter tuning was performed to obtain the optimal model parameters. The whole dataset from all types of concerned alloys was randomly split with 80% as the training set (384 data entries) and 20% left as the testing set (96 data entries). Parameter tuning was employed on the training set using grid search by five-fold cross-validation with randomly selected hyperparameters for each machine learning algorithm. The mean and variance of the five-fold MSEs were calculated to evaluate the comprehensive performance of all hyperparameters for each algorithm.

LSTM neural network

The internal structure of LSTM comprises the forget gate, input gate, and output gate. LSTM can use memory cell c_t to remove or add information. c_t is composed of the input gate i_t , forget gate f_t , and output gate o_t . First, LSTM decided what information would be discarded from the cell state by the forget gate. The previous sequence's hidden state h_{t-1} and current data X_t are simultaneously input into the forget gate f_t . The forget gate's output represented the probability of forgetting the information provided by h_{t-1} , as given in Eq. (3):

$$f_t = \sigma(W_f * [h_{t-1}, X_t] + b_f), \quad (3)$$

where W_f , b_f and σ are the linear correlation coefficient, bias, and sigmoid activation function respectively. The value of f_t is between 0 and 1, where 0 indicates that no information can pass, and 1 indicates that any information can pass.

The input gate determines how much new information is stored in the cell state, which consists of two parts, namely, i_t and \hat{c}_t , where i_t utilizes the sigmoid activation function (Eq. (4)) and \hat{c}_t utilizes the tanh activation function (Eq. (5)):

$$i_t = \sigma(W_i * [h_{t-1}, X_t] + b_i), \quad (4)$$

$$\hat{c}_t = \tanh(W_c * [h_{t-1}, X_t] + b_c), \quad (5)$$

where W_i and W_c represent linear correlation coefficients, and b_i and b_c represent biases.

The results of i_t and \hat{c}_t had to be multiplied to update the old cell state c_{t-1} to new c_t . The old state c_{t-1} was multiplied by f_t to discard the information that would certainly be discarded, and then added by the product of the input gate i_t and \hat{c}_t , as Eq. (6). For the output gate o_t (Eq. (7)), the sigmoid activation function was used and W_o and b_o indicated the correlation coefficient and bias. The hidden state h_t was finally updated by the product of o_t and c_t after the activation function \tanh (Eq. (8)), according to:

$$c_t = f_t * c_{t-1} + i_t * \hat{c}_t \quad (6)$$

$$o_t = \sigma(W_o * [h_{t-1}, X_t] + b_o), \quad (7)$$

$$h_t = o_t * \tanh(c_t). \quad (8)$$

The LSTM neural network was implemented by TensorFlow and Keras after transforming the original dataset to a set of training sequences with a sliding window, namely the lookback window. We carried out parameter tuning of the lookback window for the LSTM neural network, where we transformed the 120 days of data entries of the 4 alloys into time-series sequences with lookbacks of 3, 5, 8, and 15. For each lookback length, the LSTM neural network

was trained on the first 80% sequences as the training set and tested on the remaining 20% sequences as the testing set to evaluate the generalization ability for individual dataset related to each type of alloy. Taking Cast Stellite 6 alloy as an example, the architecture of the LSTM neural network for the cast Stellite 6 alloy consisted of an input layer with 5 units (e.g., lookback = 5), 3 hidden layers with 128 units, and a fully connected layer with 1 unit and a 'relu' activation function as the output layer. Except for the output layer, dropout was used to prevent the neural networks from overfitting. The MSE metric was applied as a loss function and the network was optimized by the 'Adam' stochastic optimization algorithm. The hyperparameters were also tuned, including the number of hidden layers and units, batch size, and dropout.

DATA AVAILABILITY

We shared the dataset to the Materials Genome Engineering Database to facilitate the data for further reuse, and the dataset can be found by this link (<https://www.mgedata.cn/search/#/153870/1064>).

CODE AVAILABILITY

The relevant code is available from the corresponding author upon reasonable request.

Received: 3 August 2022; Accepted: 30 October 2022;

Published online: 11 November 2022

REFERENCES

- Li, T., Wu, J. & Frankel, G. Localized corrosion: Passive film breakdown vs. Pit growth stability, Part VI: Pit dissolution kinetics of different alloys and a model for pitting and repassivation potentials. *Corros. Sci.* **182**, 109277 (2021).
- Williams, D., Westcott, C. & Fleischmann, M. Stochastic models of pitting corrosion of stainless steels: I. Modeling of the initiation and growth of pits at constant potential. *J. Electrochem. Soc.* **132**, 1796 (1985).
- Williams, D. E., Stewart, J. & Balkwill, P. H. The nucleation, growth and stability of micropits in stainless steel. *Corros. Sci.* **36**, 1213–1235 (1994).
- Frankel, G., Stockert, L., Hunkeler, F. & Boehni, H. Metastable pitting of stainless steel. *Corrosion* **43**, 429–436 (1987).
- Li, T. et al. Cryo-based structural characterization and growth model of salt film on metal. *Corros. Sci.* **174**, 108812 (2020).
- Li, T., Scully, J. & Frankel, G. Localized corrosion: passive film breakdown vs pit growth stability: part v. validation of a new framework for pit growth stability using one-dimensional artificial pit electrodes. *J. Electrochem. Soc.* **166**, C3341 (2019).
- Li, T., Scully, J. & Frankel, G. Localized corrosion: passive film breakdown vs. pit growth stability: part IV. The role of salt film in pit growth: a mathematical framework. *J. Electrochem. Soc.* **166**, C115 (2019).
- Newman, R. C. A comprehensive modelling and experimental approach to study the diffusion-controlled dissolution in pitting corrosion. *Corros. Sci.* **186**, 109461 (2021).
- Frankel, G. S., Li, T. & Scully, J. R. Perspective—localized corrosion: passive film breakdown vs pit growth stability. *J. Electrochem. Soc.* **164**, C180 (2017).
- Zhu, L., Zhou, J. & Sun, Z. Materials data toward machine learning: advances and challenges. *J. Phys. Chem. Lett.* **13**, 3965–3977 (2022).
- Hart, G. L. W., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).
- Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, 1–17 (2019).
- Wang, W. et al. Automated pipeline for superalloy data by text mining. *npj Comput. Mater.* **8**, 1–12 (2022).
- Liu, P. et al. Evolution analysis of γ ' precipitate coarsening in Co-based superalloys using kinetic theory and machine learning. *Acta Mater.* **235**, 118101 (2022).
- Jiang, X. et al. A strategy combining machine learning and multiscale calculation to predict tensile strength for pearlitic steel wires with industrial data. *Scr. Mater.* **186**, 272–277 (2020).
- Cai, J., Cottis, R. & Lyon, S. Phenomenological modelling of atmospheric corrosion using an artificial neural network. *Corros. Sci.* **41**, 2001–2030 (1999).

17. Parthiban, T. et al. Neural network analysis for corrosion of steel in concrete. *Corros. Sci.* **47**, 1625–1642 (2005).
18. Shen, C. et al. Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel. *Acta Mater.* **179**, 201–214 (2019).
19. Nyby, C. et al. Electrochemical metrics for corrosion resistant alloys. *Sci. Data* **8**, 1–11 (2021).
20. Taylor, C. D. & Tossey, B. M. High temperature oxidation of corrosion resistant alloys from machine learning. *npj Mater. Degrad.* **5**, 1–10 (2021).
21. Roy, A. et al. Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys. *npj Mater. Degrad.* **6**, 1–10 (2022).
22. Yan, L., Diao, Y., Lang, Z. & Gao, K. Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach. *Sci. Technol. Adv. Mat.* **21**, 359–370 (2020).
23. Diao, Y., Yan, L. & Gao, K. Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features. *Mater. Des.* **198**, 109326 (2021).
24. Zhi, Y. et al. Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model. *Corros. Sci.* **178**, 109084 (2021).
25. Pei, Z. et al. Towards understanding and prediction of atmospheric corrosion of an Fe/Cu corrosion sensor via machine learning. *Corros. Sci.* **170**, 108697 (2020).
26. Jiang, X., Yan, Y. & Su, Y. Predicting the corrosion properties of cast and hot isostatic pressed CoCrMo/W alloys in seawater by machine learning. *Anti Corros. Method. M.* **69**, 288–294 (2022).
27. Coelho, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Mater. Degrad.* **6**, 1–16 (2022).
28. Aghaaminiha, M. et al. Machine learning modeling of time-dependent corrosion rates of carbon steel in presence of corrosion inhibitors. *Corros. Sci.* **193**, 109904 (2021).
29. Shewalkar, A. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *J. Artif. Intell. Soft* **9**, 235–245 (2019).
30. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* **31**, 1235–1270 (2019).
31. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
32. Abadi, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Preprint at <https://arxiv.org/pdf/1603.04467.pdf> (2016).
33. Gulli, A. & Pal, S. *Deep learning with Keras, Vol. 2.* (Packt Publishing Ltd, 2017).
34. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
35. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems*, vol. 30 (NeurIPS Proceedings, 2017).

ACKNOWLEDGEMENTS

This work is financially supported by the National Key Research and Development Program of China (2021YFB3802101, 2021YFB3702403, 2020YFB0704503, and

2016YFB0700500), National Natural Science Foundation of China (52201061, U22A20106), USTB MatCom of Beijing Advanced Innovation Center for Materials Genome Engineering. Y.Y. is grateful to the late Prof. Anne Neville for taking him into the world of corrosion and tribology.

AUTHOR CONTRIBUTIONS

X.J. performed the dataset transformation, data analysis, and drafted the manuscript. Y.Y. performed the experiment and oversaw results, discussion, interpretation, and edited the manuscript. Y.S. provided technical expertise on machine learning, interpreted the results, and reviewed the manuscript.

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/s41529-022-00307-4>.

Correspondence and requests for materials should be addressed to Yu Yan.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022