Reviewing machine learning of corrosion prediction in a data-oriented perspective

Supplementary Information

Leonardo Bertolucci Coelho[1],, Dawei Zhang[2], Yves Van Ingelgem[1], Denis Steckelmacher[3], Ann Nowé[3], Herman Terryn[1]*

[1]Department of Materials and Chemistry, Research Group Electrochemical and Surface Engineering, Vrije Universiteit Brussel, Brussels, Belgium

[2]Beijing Advanced Innovation Center for Materials Genome Engineering, National Materials Corrosion and Protection Data Center, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, China

[3]VUB Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium

**leonardo.bertolucci.coelho@vub.be*

Supplementary Notes

The 'Machine learning for corrosion database' comprising the references [1-19] was built into a Pandas DataFrame (Python language). For the sake of representation, the entire DataFrame was segmented and is displayed hereafter. The database and associated code are available online at Mendeley Data and GitHub (Jupyter Notebook files).

Different metrics are possible to evaluate the prediction accuracy of regression models. However, only papers providing relative metrics (MAPE, R^2) were included in this database. We tried as much as possible to include descriptors of all major ML procedure steps, including data collection ('Data acquisition"), data cleaning feature engineering ('Feature reduction'), model validation ('Train-Test split'*), etc.

The target property of corrosion prediction varies and is usually described by many descriptors. For example, the corrosion potential '*E* (mV)' might be an input ('Selected features') and/or output ('Targets'). The 'Feature importance' column was used when formal feature importance methods were employed. Regarding the 'Ensemble' type of models ('Type of ML model' column), these do not refer to combinations of models with feature engineering (for example, "PCA-RF" was considered as a "Tree-based" type of model, and not an 'Ensemble'). For a summarized description of the ML models here discussed, the reader should refer to the 'Models description' feature.

For plotting the Fig. 7 of the Review article, the following considerations were applied to the 'Targets' attribute: when the corrosion rate 'CR' presented 'μA/cm²' as a unit, it was considered as '*I*' (current); 'material loss' represents 'weight loss' or 'volume loss'; 'depth' stands for 'defect growth' or 'corrosion depth'; and 'crack growth rate' was considered as 'CR'. The last columns of the DataFrame condensate the gathered knowledge from the references in the following form: 'Premises/challenges', 'Achievements', 'Contributions', and 'Critical points'.

*the total dataset is typically split into training sets and testing (unknown data) sets for performance evaluation of the model. Nonetheless, sometimes only the training or the testing performances were reported ('?' marks were added in the respective evaluation metric field(s)). The 'Average R^2' was sometimes considered for studies employing 'CV' (cross-validation) on the dataset. For a detailed description of the ML basic procedures, the reader could refer to the References topic in the Review article.

The processed data required to reproduce these findings are available to download from Bertolucci Coelho, Leonardo (2021), 'Machine learning for corrosion database', Mendeley Data, V1, doi:

10.17632/jfn8yhrphd.1, permanently stored at the repository via the link https://data.mendeley.com/datasets/jfn8yhrphd/1.

Supplementary Tables

Supplementary table 1. The 'Machine learning for corrosion database' in a Pandas DataFrame. The ML works from references [1-19] are organized in 34 rows and described across 49 columns.

Reference	DOI	Year	Corrosion topic	Orientation strategy	Material
1	10.1016/j.matdes.2020.109326	2021	Atmospheric corrosion	Material & Environment	LAS (20)
1	10.1016/j.matdes.2020.109326	2021	Atmospheric corrosion	Material & Environment	LAS (20)
1	10.1016/j.matdes.2020.109326	2021	Atmospheric corrosion	Material & Environment	LAS (20)
1	10.1016/j.matdes.2020.109326	2021	Atmospheric corrosion	Material & Environment	LAS (20)
2	10.1080/14686996.2020.1746196	2020	Atmospheric corrosion	Material & Environment	LAS (18)
3	10.3390/met9030383	2019	Atmospheric corrosion	Material & Environment	LAS (17)
4	10.1016/j.corsci.2020.108697	2020	Atmospheric corrosion	Environment & Instrument	steel/copper galvanic couple
4	10.1016/j.corsci.2020.108697	2020	Atmospheric corrosion	Environment & Instrument	steel/copper galvanic couple
5	10.1108/ACMM-11-2017-1858	2017	Atmospheric corrosion	Material, System & Environment	carbon steels (6)
6	10.1016/j.corsci.2009.10.024	2010	Atmospheric corrosion	Material & Environment	carbon and alloy steels
7	10.1016/j.corsci.2010.11.028	2011	Corrosion-resistant alloys	Material & Environment	Alloy 22
8	10.1016/j.corsci.2008.10.038	2009	Marine Corrosion	System & Environment	3C steel
9	10.1016/j.jmst.2020.01.044	2020	Atmospheric corrosion	Material & Environment	LAS (17)
10	10.3390/bdcc3020028	2019	Pipeline corrosion	System & Environment	X52 grade pipeline
10	10.3390/bdcc3020028	2019	Pipeline corrosion	System & Environment	X52 grade pipeline
11	10.1016/S0010-938X(99)00024-4	1999	Atmospheric corrosion	Environment	steel
11	10.1016/S0010-938X(99)00024-4	1999	Atmospheric corrosion	Environment	Zn
12	10.1108/00035591111148894	2011	Rebar corrosion / pitting	Environment	rebar steel (ASTM A615 steel)
13	10.1109/OCEANS-Genova.2015.7271592	2015	Pipeline corrosion	System	pipeline (?)
13	10.1109/OCEANS-Genova.2015.7271592	2015	Pipeline corrosion	System	pipeline (?)
14	10.1016/j.measurement.2020.108141	2020	Rebar corrosion	Material	steel bar
14	10.1016/j.measurement.2020.108141	2020	Rebar corrosion	Material	steel bar
6	10.1016/j.corsci.2009.10.024	2010	Corrosion-resistant alloys	Material, System & Environment	metallic glasses
6	10.1016/j.corsci.2009.10.024	2010	Crevice corrosion	System & Instrument	Ti-2
7	10.1016/j.corsci.2010.11.028	2011	Crevice corrosion	Material, System & Environment	Alloy 22
15	10.1016/j.corsci.2014.08.011	2014	Stress corrosion cracking	Material, System & Environment	304 SS
16	10.1016/j.corsci.2016.02.008	2016	Inhibitors	Material & Environment	AA7075 (pH 4)
16	10.1016/j.corsci.2016.02.008	2016	Inhibitors	Material & Environment	AA7075 (pH 10)
16	10.1016/j.corsci.2016.02.008	2016	Inhibitors	Material & Environment	AA2024 (pH 4)
16	10.1016/j.corsci.2016.02.008	2016	Inhibitors	Material & Environment	AA2024 (pH 10)
17	10.1016/j.corsci.2019.01.046	2019	Conversion coatings	Design-oriented	AZ91D Mg
18	10.1016/j.engappai.2016.09.008	2017	Rebar corrosion / pitting	Environment	rebar steel (ASTM A615 steel)
18	10.1016/j.engappai.2016.09.008	2017	Marine corrosion	System & Environment	3C steel
19	10.1016/j.corsci.2021.109438	2021	Rebar corrosion / pitting	Material, System & Environment	carbon steel

Reference	Data nature	Data collection	Data acquisition
1	Experimental	in-field (4)	weight loss & Meteorology
1	Experimental	in-field (4)	weight loss & Meteorology
1	Experimental	in-field (4)	weight loss & Meteorology
1	Experimental	in-field (4)	weight loss & Meteorology
2	Experimental	in-field (3)	weight loss & Meteorology
3	Experimental	in-field (6)	weight loss & Meteorology
4	Experimental	in-field (1)	current sensor & Meteorology
4	Experimental	in-field (1)	current sensor & Meteorology
5	Experimental	in-field (2)	weight loss
6	Experimental	in-field (literature)	NIST data
7	Experimental	lab (literature (1))	weight loss (Basic Saturated Water)
8	Experimental	lab (literature: https://www.jcscp.org/EN/abstract/abstract2324.shtml)	electrochemical technique (seawater environment)
9	Experimental	in-field (6) (literature)	weight loss & Meteorology
10	Experimental	in-field (60) (literature)	routine monitoring / simulation
10	Simulation	data generation	multivariate polynomial regression
11	Experimental	in-field (literature (42))	annual averages, climatological tables
11	Experimental	in-field (literature (42))	annual averages, climatological tables
12	Experimental	lab	polarization measurements (simulated concrete pore solutions)
13	Experimental & simulation	in-field (2) & multiphase flow simulation (OLGA software)	geometrical pipeline characteristics, flow simulation, deterministic models
13	Experimental & simulation	in-field (2) & multiphase flow simulation (OLGA software)	geometrical pipeline characteristics, flow simulation, deterministic models
14	Experimental	lab	OCP (5% NaCl)
14	Experimental	lab	OCP (5% NaCl)
6	Experimental	lab (literature (1))	polarization measurements
6	Experimental	lab (literature)	electrochemical cell (0.27 M NaCl
7	Experimental	lab (literature (1))	polarization measurements (chloride media)
15	Experimental	lab (literature (30))	various
16	Experimental & simulation	lab & generation of molecules (Sybyl x2.0 molecular modelling (Certera ltd))	image analysis from high-throughput & simulation of chemical/molecular properties
16	Experimental & simulation	lab & generation of molecules (Sybyl x2.0 molecular modelling (Certera ltd))	image analysis from high-throughput & simulation of chemical/molecular properties
16	Experimental & simulation	lab & generation of molecules (Sybyl x2.0 molecular modelling (Certera ltd))	image analysis from high-throughput & simulation of chemical/molecular properties
16	Experimental & simulation	lab & generation of molecules (Sybyl x2.0 molecular modelling (Certera ltd))	image analysis from high-throughput & simulation of chemical/molecular properties
17	Experimental	lab	hydrogen evolution test
18	Experimental	lab (literature: DOI=10.1108/00035591111148894)	polarization measurements (simulated concrete pore solutions)
18	Experimental	lab (literature: https://www.jcscp.org/EN/abstract/abstract2324.shtml)	electrochemical technique (seawater environment)
19	Experimental	variable (literature: https://www.sciencedirect.com/science/article/pii/S0010938X21002055)	various

Reference	Data availability	Dataset size
1	http://data.ecorr.org/	85.0
1	http://data.ecorr.org/	85.0
1	http://data.ecorr.org/	85.0
1	http://data.ecorr.org/	85.0
2	https://smds.nims.go.jp/corrosion/en/	306.0
3	http://data.ecorr.org/	409.0
4		48647.0
4		48647.0
5	http://data.ecorr.org/	6.0
6	https://www.astm.org/DIGITAL_LIBRARY/STP/PAGES/STP13769S.htm	NaN
7	https://doi.org/10.5006/1.3287856	NaN
8	http://dx.doi.org/10.1016/j.engappai.2016.09.008	46.0
9	https://www.mdpi.com/2075-4701/9/3/383	409.0
10	iopscience.iop.org/article/10.1149/2.0701506jes/pdf, https://doi.org/10.1016/j.engfailanal.2015.11.052	8300.0
10	iopscience.iop.org/article/10.1149/2.0701506jes/pdf, https://doi.org/10.1016/j.engfailanal.2015.11.052	5000.0
11	Appendix A	406.0
11	Appendix C	315.0
12	Table 1	62.0
13		NaN
13		NaN
14		80.0
14		80.0
6	https://link.springer.com/article/10.1023/A:1018518704635	300.0
6	https://www.sciencedirect.com/science/article/abs/pii/S0010938X04002136	8.0
7	Table 2	9.0
15	various references	280.0
16	http://dx.doi.org/10.1016/j.corsci.2016.02.008	100.0
16	http://dx.doi.org/10.1016/j.corsci.2016.02.008	100.0
16	http://dx.doi.org/10.1016/j.corsci.2016.02.008	100.0
16	http://dx.doi.org/10.1016/j.corsci.2016.02.008	100.0
17	Appendix 2	36.0
18	Table A1	62.0
18	Table B1	46.0
19		1242.0

Reference	Initial N° of attributes	Attributes interdependence	Feature reduction	Feature creation
1	26		GBDT+Kendall	False
1	26		PCA	False
1	26		GBDT+Kendall	method I
1	26		GBDT+Kendall	method II
2	17	positive PCC for all input attributes analysed	PCC+MIC	False
3	15		False	False
4	10		False	False
4	10		False	False
5	7		False	False
6	17		False	False
7	4		False	False
8	6		False	False
9	16		False	False
10	14		False	False
10	14		PCA	False
11	6		False	False
11	6		False	False
12	5		False	False
13	16		False	False
13	16		False	False
14	7		False	False
14	7		False	False
6	11		False	False
6	12		False	False
7	7		False	False
15	7		False	False
16	20		sparse feature selection (BioModeller)	molecular descriptors (DRAGON program)
16	14		sparse feature selection (BioModeller)	molecular descriptors (DRAGON program)
16	32		sparse feature selection (BioModeller)	molecular descriptors (DRAGON program)
16	34		sparse feature selection (BioModeller)	molecular descriptors (DRAGON program)
17	4		False	False
18	5		False	False
18	6		False	False
19	18	ECP, Eb = f(T, pH); pH = f(T, cement composition, water/cement ratio, porosity)	False	False

Reference	Selected features	Time selected?
1	O2_max, S_max, pH_min, T_mean, Cr, C, Si, P, S, t	True
1	9 components	False
1	O2_max, S_max, pH_min, T_mean, t, ENs, K, Rm, Rc	True
1	O2_max, S_max, pH_min, T_mean, t, lambda-, V-, ENp-, deltaHf-	True
2	elements content, T_MAX, T_AVE, RH_MIN, PRECIPIT, SOLAR, Cl-, SO2, t	True
3	RH_mean, T_Mean, Rainfall, SO2, pH of rain, Cl-, Mn, S, P, Si, Cr, Cu, Ni, t	True
4	T, RH, rainfall	False
4	T, RH, rainfall, QACM	False
5	station, steel, CR-year1, CR-y2, CR-y4, CR-y8	False
6	B, C, Cr, Cu, Fe, Mn, Mo, Ni, P, S, Si, Ti, V, duration of exposure, log(duration of exposure)	True
7	temperature, treatment (welded or annealed), exposure time)	True
8	temperature, dissolved oxygen, salinity, pH, E (mV)	False
9	RH_mean, T_Mean, Rainfall, SO2, pH of rain, Cl, Mn, S, P, Si, Cr, Cu, Ni, Fe, t	True
10	TM, PS, PCO2, pH, Cl, SO4, BSW, BOPD, BWPD, MMCFD, Fe, HCO3, Ca	False
10	TM, PS, PCO2, pH, Cl, SO4, BSW, BOPD, BWPD, MMCFD, Fe, HCO3, Ca	False
11	Temperature , TOW, SO2, Cl, Exposure time	True
11	Temperature , TOW, SO2, Cl, Exposure time	True
12	CITT, chloride binding, pH, DO	False
13	elevat, inclinat, concav, concav/inclinat, dist, flow, hold-up, pres, gas flow, tot flow, liq veloc, gas veloc, de Waard or NORSOK	False
13	elevat, inclinat, concav, concav/inclinat, dist, flow, hold-up, pres, gas flow, tot flow, liq veloc, gas veloc, de Waard or NORSOK	False
14	cement (CE), limestone powder (LSP), coarse aggregate (CA), fine aggregate (FA), water (WA), and exposure period (EXP)	True
14	cement (CE), limestone powder (LSP), coarse aggregate (CA), fine aggregate (FA), water (WA), and exposure period (EXP)	True
6	Mo (x), Ni (14 - x), (x + 1), HCl, H2SO4, E (mV), dE	False
6	incubation period, duration	True
7	temperature, treatment (welded or annealed), pH, Cl-, aging time, type (general or localized)	False
15	E (V), conductivity, degree of sensitization (DoS), stress intensity factor (KI), T, pH	False
16	alloy, pH, 17 descriptors	False
16	alloy, pH, 11 descriptors	False
16	alloy, pH, 11-29 descriptors	False
16	alloy, pH, 14-31 descriptors	False
17	pH, Total Acidity, TA/pH	False
18	total Cl- (Cl TT), Cl- binding, pH, DO	False
18	Temp, DO, salinity, pH, E (mV)	False
19	pH, T, ECP (0.5 ppm, 1 ppm, 5 ppm, 8 ppm [O2]), Eb, porosity, w/cem ratio, SiO2, Al2O3, Fe2O3, CaO, MgO, Na2O, K2O, SO3	False

Reference	Data dimension	Targets	Feature importance
1	10	CR (mm/a)	GBDT
1	9	CR (mm/a)	GBDT
1	9	CR (mm/a)	GBDT
1	9	CR (mm/a)	GBDT
2	9	CR (mm/a)	RF
3	14	CR (μm/a)	RF
4	3	I (nA)	RF
4	4	I (nA)	RF
5	6	CR (μm/a)	False
6	15	CR (μm/a)	False
7	3	weight loss (mg)	False
8	5	CR ($\mu A/cm^2$)	False
9	15	CR ($\mu A/cm^2$)	DCCF-WKNNs
10	13	corrosion defect depth growth (mm/a)	False
10	13	corrosion defect depth growth (mm/a)	False
11	5	corrosion depth (mm)	False
11	5	corrosion depth (mm)	False
12	4	pitting risk (Epit-Ecorr) (mV)	False
13	14	CR (mm/a)	False
13	14	volume loss (mm^3)	False
14	6	E (mV)	False
14	6	E (mV)	False
6	7	((log(i), icorr (HCl)), (log(i), icorr (H2SO4)), Ecorr (HCl), Ecorr (H2SO4), Rp (HCl), Rp (H2SO4))	False
6	2	(Max penetration depth (mm), Weight change (g), Q (C), O2 consumed (%))	False
7	6	crevice repassivation E (mV)	ANN weighting
15	6	crack growth rate (cm/s)	Sensitivity analysis
16	19	inhibition score (0-10)	False
16	13	inhibition score (0-10)	False
16	(13, 31)	inhibition score (0-10)	False
16	(16, 33)	inhibition score (0-10)	False
17	3	HER (mL/cm.h)	False
18	4	Pitting risk (Ecorr - Epit) (mV)	False
18	5	CR ($\mu A/cm^2$)	False
19	17	CT (%Total Cl/cem)	False

Reference	Knowledge mining	Train-Test split	Training split
1	False	(80, 20)	LOOCV
1	False	(80, 20)	LOOCV
1	False	(80, 20)	LOOCV
1	False	(80, 20)	LOOCV
2	False	(80, 20)	5-fold CV
3	RF	(66.5, 33.5)	False
4	False	(90, 10)	False
4	False	(90, 10)	False
5	True	(1, 5)	(False, or LOOCV)
6	False	True	False
7	False	CV	False
8	False	((41, 5), (42, 4), LOOCV)	False
9	DCCF-WKNNs	True	(True, for DCCF-WKNNs)
10	False	False	False
10	False	5-fold CV	False
11	sensitivity analysis	(90, 10)	False
11	sensitivity analysis	(90, 10)	False
12	True	(90, 10)	False
13	False	?	?
13	False	?	?
14	False	10-fold CV	False
14	False	((67, 33), (70, 30), (80, 20), (85, 15))	False
6	False	(159, 141)	?
6	False	False	False
7	False	(2-fold CV, 80, 20)	False
15	False	(85, 15)	(70, 15)
16	False	(80, 20)	False
16	False	(80, 20)	False
16	False	(80, 20)	False
16	False	(80, 20)	False
17	False	(30, 6)	False
18	False	(90, 10)	False
18	False	(90, 10)	False
19	False	(85, 15)	(70, 15)

Reference	ML models	Type of ML model
1	RF	Tree-based
1	RF	Tree-based
1	RF	Tree-based
1	RF	Tree-based
2	(MLR, RR, SVR, RF, GBDT, XGBoost)	(Linear, Linear, Kernel-based, Tree-based, Tree-based, Ensemble)
3	(ANN, SVR, LR, RF)	(ANN, Kernel-based, Linear, Tree-based)
4	(ANN, SVR, RF)	(ANN, Kernel-based, Tree-based)
4	(ANN, SVR, RF)	(ANN, Kernel-based, Tree-based)
5	(GM(1,1), GANGBM(1,1), RGANGBM(1,1), BPNN, SVR)	(Gray, Genetic, Genetic, ANN, Kernel-based)
6	ANN	ANN
7	ANN	ANN
8	(BPNN, BPNN, SVR-PSO, SVR-PSO, SVR-PSO)	(ANN, ANN, Ensemble, Ensemble, Ensemble)
9	(ANN, SVR, RF, RF-WKNNs, cForest, DCCF-WKNNs)	(ANN, Kernel-based, Tree-based, Ensemble, Tree-based, Ensemble)
10	multivariate polynomial regression	Linear
10	(PSO-FFANN, GBM, RF, DNN, PCA-PSO-FFANN, PCA-GBM, PCA-RF, PCA-DNN)	(Ensemble, Tree-based, Tree-based, ANN, Ensemble, Tree-based, Tree-based, ANN)
11	ANN	ANN
11	ANN	ANN
12	BP NN	ANN
13	(ANN, ensemble ANN)	(ANN, ANN)
13	(ANN, ensemble ANN)	(ANN, ANN)
14	(RF, LR, FNN, SVR, ANN)	(Tree-based, Linear, ANN, Kernel-based, KNN)
14	RF	Tree-based
6	ANN	ANN
6	ANN	ANN
7	ANN	ANN
15	ANN	ANN
16	(MLREM, BRANNGP, BRANNLP)	(Linear, Ensemble, Ensemble)
16	(MLREM, BRANNGP, BRANNLP)	(Linear, Ensemble, Ensemble)
16	(MLREM, BRANNGP, BRANNLP, MLREM, BRANNGP)	(Linear, Ensemble, Ensemble, Linear, Ensemble)
16	(MLREM, BRANNGP, MLREM, BRANNGP)	(Linear, Ensemble, Linear, Ensemble)
17	ANN	ANN
18	(SVR, ANNs+SVR+LR, SVR ensemble, ANNs+CART+LR, SVMs+LR, SVMs+CART, SFA-LSSVR)	(Kernel-based, Ensemble, Kernel-based, Ensemble, Ensemble, Ensemble, Ensemble)
18	(SVR, CART+LR, LR, CART+SVR+LR, SVMs+SVR, SVMs+CART, SFA-LSSVR)	(Kernel-based, Ensemble, Linear, Ensemble, Kernel-based, Ensemble, Ensemble)
19	KSOM-ANN	Ensemble

Reference	New models proposed	Models description
1	False	
1	False	
1	False	
1	False	
2	False	
3	False	
4	False	
4	False	Model with Q (charge) correction
5	True	(GM(1,1)=Grey model, GANGBM(1,1)=genetic algorithm non-linear gray Bernoulli model, RGANGBM(1,1)=regularization genetic algorithm non-linear gray Bernoulli model)
6	False	
7	False	
8	False	BPNN (41 training), BPNN (42 training), SVR (41 training), SVR (42 training), SVR (LOOCV)
9	True	RF-WKNNs=random forests-weighted k-nearest neighbors. DCCF-WKNNs (Densely connected cascade forests-WKNNs): combination of cascade forest and RF-WKNNs (layer-by-layer representation method). 4 layers (DCCF-WKNNs) were used (1 layer=4 RF-WKNNs)
10	False	
10	False	(Gradient Boosting Machine=GBM, DNN=Deep Neural Network)
11	False	
11	False	
12	False	
13	False	
13	False	
14	False	
14	False	RF with 4 different Train Test splits
6	False	BPNN with 6 different Targets
6	False	BPNN with 4 different Targets
7	False	
15	False	
16	False	MLREM=multiple linear regressions; feed forward Bayesian regularized neural networks using either a Gaussian prior (BRANNGP) or sparsity-inducing Laplacian prior (BRANNLP)
16	False	MLREM=multiple linear regressions; feed forward Bayesian regularized neural networks using either a Gaussian prior (BRANNGP) or sparsity-inducing Laplacian prior (BRANNLP)
16	False	MLREM (11 descriptors), BRANNGP (11 d), BRANNLP (11 d), MLREM (29 d), BRANNGP (29 d)
16	False	MLREM (14 descriptors), BRANNGP (14 d), MLREM (31 d), BRANNGP (31 d)
17	False	BPNN (pH), BPNN (TA), BPNN (TA/pH)
18	False	SFA-LSSVR=hybrid metaheuristic regression model integrating the nature-inspired metaheuristic optimization (smart firefly algorithm (SFA)) and least squares support vector regression (LSSVR)
18	False	SFA-LSSVR=hybrid metaheuristic regression model integrating the nature-inspired metaheuristic optimization (smart firefly algorithm (SFA)) and least squares support vector regression (LSSVR)
19	False	Kohonen-self organized map (KSOM) and ANN coupled methodology was developed to find missing values and for regression

Reference	Non traditional parameters
1	
1	
1	
1	
2	
3	
4	
4	
5	RGANGBM(1,1): the magnitude of eta is related to the influence of the regularization term, and is determined by the curve trend of historical data of series-data
6	
7	
8	
9	RF-WKNNs: 2 additional parameters based on RF. The first parameter is a threshold, the second is the K of KNNs. cForest: the combination of several layers, each layer contains several RFs. DCCF-WKNNs: same parameters as in RF-WKNNs and cForest
10	
10	
11	
11	
12	
13	
13	
14	
14	
6	
6	
7	
15	
16	
16	
16	
16	
17	
18	
18	
19	

Reference	Software	Libraries	Evaluation method
1	Python	Scikit-learn	R^2
1	Python	Scikit-learn	R^2
1	Python	Scikit-learn	R^2
1	Python	Scikit-learn	R^2
2	Python	Scikit-learn	R^2
3	Python	Scikit-learn	R^2, MAPE
4	Python	Scikit-learn	R^2
4	Python	Scikit-learn	R^2
5			MAPE
6	NeuralWare		R^2
7	NeuralWare		R^2
8			R^2, MAPE
9	Python	Scikit-learn	R^2, MAPE
10			R^2
10			MAPE
11	Commercial software		R, MAPE
11	Commercial software		R, MAPE
12			R^2
13			R
13			R
14			R
14			R
6	NeuralWare		R^2
6	NeuralWare		R^2
7	NeuralWare		R^2
15	MATALAB		R^2
16		ars.els-cdn.com/content/image/1-s2.0-S0010938X12000509-mmc1.zip	R^2
16		ars.els-cdn.com/content/image/1-s2.0-S0010938X12000509-mmc1.zip	R^2
16		ars.els-cdn.com/content/image/1-s2.0-S0010938X12000509-mmc1.zip	R^2
16		ars.els-cdn.com/content/image/1-s2.0-S0010938X12000509-mmc1.zip	R^2
17	MATALAB		R^2
18	MATALAB, WEKA		MAPE
18	MATALAB, WEKA		MAPE
19	MATALAB, RStudio	Kohonen, CPANN toolboxes (MATLAB), Kohonen (RStudio)	R^2

	Metric conversions	R (train)	R (test)
Reference			
1		None	None
1		None	None
1		None	None
1		None	None
2		None	None
3		None	None
4		None	None
4		None	None
5		None	None
6		None	None
7		None	None
8		None	None
9		None	None
10		None	None
10		None	None
11	R to R^2	None	[0.74]
11	R to R^2	None	[0.78]
12		None	None
13	R to R^2	[0.83, 0.94]	None
13	R to R^2	[0.58, 0.82]	None
14	R to R^2	None	[0.947, 0.922, 0.955, 0.934, 0.937]
14	R to R^2	[0.993, 0.993, 0.993, 0.99]	[0.968, 0.967, 0.974, 0.987]
6		None	None
6		None	None
7		None	None
15		None	None
16		None	None
16		None	None
16		None	None
16		None	None
17		None	None
18		None	None
18		None	None
19		None	None

Reference	R^2 (train)	R^2 (test)	Average R^2
1	0.94	0.92	
1	0.87	0.75	
1	0.89	0.7	
1	0.91	0.9	
2	(0.54, 0.56, 0.9, 0.94, 0.96, 0.93)	(0.43, 0.33, 0.48, 0.73, 0.69, 0.77)	
3	(0.858, 0.404, 0.623, 0.936)	(0.781, 0.432, 0.505, 0.891)	
4	(0.027, 0.211, 0.526)	(0.026, 0.226, 0.457)	
4	(0.202, 0.898, 0.94)	(0.025, 0.91, 0.94)	
5	None	None	
6	None	0.987	
7	None	None	0.8
8	None	(0.91, 0.931, 0.942, 0.968, 0.836)	
9	(1.0, 0.96, 0.98, 0.98, 1.0, 1.0)	(0.785, 0.823, 0.908, 0.911, 0.92, 0.924)	
10	0.9819	None	
10	None	None	
11	None	[0.548]	
11	None	[0.608]	
12	0.8741	None	
13	[0.689, 0.884]	None	
13	[0.336, 0.672]	None	
14	None	[0.897, 0.85, 0.912, 0.872, 0.878]	
14	[0.986, 0.986, 0.986, 0.98]	[0.937, 0.935, 0.949, 0.974]	
6	None	None	(1, 1, 1, 1, 0.87, 0.89)
6	(1, 1, 1, 1)	None	
7	None	None	1
15	0.95	None	
16	(0.77, 0.71, 0.71)	(0.79, 0.82, 0.79)	
16	(0.67, 0.65, 0.64)	(0.56, 0.6, 0.58)	
16	(0.48, 0.4, 0.32, 0.85, 0.75)	(0.62, 0.63, 0.65, 0.75, 0.82)	
16	(0.62, 0.52, 0.88, 0.81)	(0.46, 0.48, 0.76, 0.74)	
17	None	None	(0.19, 0.49, 0.86)
18	None	None	
18	None	None	
19	0.99	None	

Reference	MAPE (%) (train)	MAPE (%) (test)
1		
1		
1		
1		
2		
3	(18.78, 40.81, 42.36, 11.32)	(22.12, 41.35, 45.63, 16.08)
4		
4		
5		
6		
7		
8		(5.001, 4.215, 3.84, 3.18, 1.36)
9	(0.28, 3.66, 6.02, 6.02, 0.78, 0.89)	(28.16, 25.16, 16.21, 15.31, 15.22, 12.95)
10		
10		(34.1329, 31.9266, 32.4267, 23.647, 7.8588, 6.0082, 7.7421, 6.6813)
11		39
11		53
12		
13		
13		
14		
14		
6		
6		
7		
15		
16		
16		
16		
16		
17		
18	(131.53, 145.52, 124.55, 152.71, 142.92, 57.72, 0.0)	(155.8, 169.07, 157.59, 162.94, 159.6, 59.49, 5.6)
18	(2.94, 13.51, 13.95, 17.08, 14.24, 9.87, 0.18)	(14.56, 14.15, 15.39, 17.18, 14.96, 9.68, 1.26)
19	15	

Reference	Sampling periodicity (years)	Sub-datasets with different times	R^2 (time subsets) (validation)
1	(1, 2)	False	
1	(1, 2)	False	
1	(1, 2)	False	
1	(1, 2)	False	
2	(1, 2, 3, 5, 7, 10)	True	
3	(1, 2, 4, 8, 16)	True	(, , , (0.898, 0.878, 0.873, 0.866, 0.733))
4	0.09	True	
4	0.09	True	
5	(1, 2, 4, 8, 16)	True	
6	?	False	
7	?	True	0.8
8	?	False	
9	(1, 2, 4, 8, 16)	True	
10	?	False	
10		False	
11	?	False	
11	?	False	
12		False	
13		False	
13		False	
14		False	
14		False	
6		False	
6	?	True	
7		False	
15		False	
16	?	False	
16	?	False	
16	?	False	
16	?	False	
17		False	
18		False	
18		False	
19		False	

Reference	R^2 (time subsets) (test)	MAPE (time subsets) (validation))	MAPE (time subsets) (test))
1			
1			
1			
1			
2			
3	(, , , (0.898, 0.878, 0.873, 0.866, 0.733))	(, , , (10.06, 10.18, 9.45, 8.98, 18.0))	
4			
4			
5			((0, 3, 9, 13, 29.25), (0, 1, 3, 3, 16.77), (0, 2, 5, 6, 9.15), (, , , , 28.02), (, , , , 83.55))
6			
7			
8			
9	article Fig. 5		article Fig. 5
10			
10			
11			
11			
12			
13			
13			
14			
14			
6			
6			
7			
15			
16			
16			
16			
16			
17			
18			
18			
19			

Reference	Premises/challenges
1	Application of CR prediction models is often limited to alloys with similar composition with the ones used for training (inputs with fixed number of elements content)
1	Application of CR prediction models is often limited to alloys with similar composition with the ones used for training (inputs with fixed number of elements content)
1	Application of CR prediction models is often limited to alloys with similar composition with the ones used for training (inputs with fixed number of elements content)
1	Application of CR prediction models is often limited to alloys with similar composition with the ones used for training (inputs with fixed number of elements content)
2	The formation of rust layers on steel surfaces impacts the CR as a function of their exposure periods
3	The three main challenges with atmospheric corrosion datasets from multiple materials & environments: (1) non-linearity, (2) small samples, (3) steep-manifold structure
4	Traditional methods to predict atmospheric corrosion requires the accumulation of environmental data on a yearly basis. These yearly averages could generate inaccurate predictions especially in the short-term
4	Traditional methods to predict atmospheric corrosion requires the accumulation of environmental data on a yearly basis. These yearly averages could generate inaccurate predictions especially in the short-term
5	GANGBM(1,1) have been used to model the small sample datasets with high performance. Nevertheless, when applied to the CR series data, an overfitting phenomenon occurs: the prediction accuracy is high only for historical data, but not for future data
6	Among the most important challenges in corrosion studies of certain metals and alloys is to learn the trends that govern the long term corrosion behavior. The environmental effects on the behavior are sometimes not well understood
7	The prediction of the corrosion behaviour based on weight loss measurements takes considerable time, and is particularly difficult in the case of Alloy 22, which is very corrosion resistant
8	ANN was proven to be effective for predicting the corrosion performance of carbon steel, however, it has weak generalization ability for small-sample dataset, and is prone to over-fitting
9	The corrosion predictions of various materials in various outdoor environments still pose challenges to commonly used models (ANN, SVM, power-linear function, hidden Markov) applied to corrosion indoor tests with just one single material
10	The first step in the Corrosion Risk Assessment model development is to determine the requisite input parameters for estimating the corrosion defect depth amongst many parameters routinely measured by the oil and gas companies
10	The first step in the Corrosion Risk Assessment model development is to determine the requisite input parameters for estimating the corrosion defect depth amongst many parameters routinely measured by the oil and gas companies
11	In contrast to previous works using linear regression technique, the ANN is promising for modelling non-linear/complex systems based on interpolation from past experience
11	In contrast to previous works using linear regression technique, the ANN is promising for modelling non-linear/complex systems based on interpolation from past experience
12	The lack of understanding of the Cl threshold value of pitting corrosion of steel in concrete is attributable to its numerous affecting factors, such as: the pH, the E of the steel, and the physical condition of the steel/concrete interface
13	The integration of the geometrical profile of a real pipeline, flow simulations and deterministic models could strongly improve the predictive performance
13	The integration of the geometrical profile of a real pipeline, flow simulations and deterministic models could strongly improve the predictive performance
14	The high nonlinearity nature of embedded steel corrosion due to complex features of reinforced concrete structures makes corrosion properties prediction difficult due to the lack of theoretical basis for some phenomena
14	The high nonlinearity nature of embedded steel corrosion due to complex features of reinforced concrete structures makes corrosion properties prediction difficult due to the lack of theoretical basis for some phenomena
6	Data mining short-cuts the deterministic model and provides a hidden map between measured alloy characteristics, environments, and the experimental failure observations on alloys
6	Among the most important challenges in corrosion studies of certain metals and alloys is to learn the trends that govern the long term corrosion behavior
7	The application of ANN is limited in corrosion studies of Alloy 22, which is very corrosion resistant
15	The CGR data from sensitized stainless, when estimated by modelling, often present a dispersion of more than two orders of magnitude
16	A novel high throughput corrosion inhibition testing experiments was combined with ML to model and predict the corrosion inhibition performance of a large library (100) of organic inhibitor candidates
16	A novel high throughput corrosion inhibition testing experiments was combined with ML to model and predict the corrosion inhibition performance of a large library (100) of organic inhibitor candidates
16	A novel high throughput corrosion inhibition testing experiments was combined with ML to model and predict the corrosion inhibition performance of a large library (100) of organic inhibitor candidates
16	A novel high throughput corrosion inhibition testing experiments was combined with ML to model and predict the corrosion inhibition performance of a large library (100) of organic inhibitor candidates
17	A new designing strategy, i.e. the ratio of the total acidity (TA) and pH value (TA/pH) was proposed to validate the chemical formulation of a phosphate-based conversion coating
18	Various ML models are typically used to construct single classifiers, but ensemble systems compensates for errors made by the individual classifiers. Nature-inspired metaheuristic optimization algorithms such as FA can effectively solve optimization problems
18	Various ML models are typically used to construct single classifiers, but ensemble systems compensates for errors made by the individual classifiers. Nature-inspired metaheuristic optimization algorithms such as FA can effectively solve optimization problems
19	Active corrosion of carbon steel in reinforced concrete occurs when [Cl] exceeds the Cl threshold. CT is affected by many physical/chemical parameters and has been the subject of numerous investigations (existance of considerable data)

Reference	Achievements
1	The GBDT + Kendall feature reduction method provided better predictive accuracies than PCA
1	The GBDT + Kendall feature reduction method provided better predictive accuracies than PCA
1	Through feature creation methods based on physical properties simulating elements, the model's applicability can be expanded to other alloys
1	Through feature creation methods based on physical properties simulating elements, the model's applicability can be expanded to other alloys
2	Based on domain knowledge, the SO2 feature was selected. RF, GBDT and XGBoost models showed high predictive accuracy
3	The poor performance of SVR was related to the steep-manifold structure of the data (specially in case of small number of samples). The unstable classifiers (RF, ANN) produced better performances than the other models. In contrast to ANN, the RF performance was less affected by the small number of training samples
4	First results indicated that as the time increased the models lost accuracy
4	By considering the rust formation on the sensor (QACM) as an input parameter, the prediction accuracy improved for all models (QACM was the most important factor in the second half of the test). RF with charge correction produced a high-accuracy model
5	The RGANGBM(1,1) model was proposed to realize the long-term (Year 16) prediction on CR vs time series. A regularization term eta was added to genetic algorithm to implement parameter optimization of NGBM (1,1) and to avoid the overfitting problem. Despite using one single training data, the prediction accuracy of the proposed method was relatively high
6	The ANN model presented a high prediction accuracy for long time periods, despite being trained with short time period data only
7	More data points would be valuable to test the robustness of the BPNN model
8	The higher the number of training samples, the better was the regression performance of the BPNN or SVR models. The generalization ability of SVR was superior to that of BPNN
9	The new deep-structure model (DCCF-WKNNs) provided a useful tool for the modelling and knowledge-mining of LAS corrosion in outdoor atmospheric (high generalization ability)
10	Data bias was avoided by the generation of datasets (multivariate polynomial regression) comprising a mixture of low, mild, high and severe corrosion defect depth growth rates
10	Improvement in the accuracy with PCA transformations highlighted the potency of ML for the prediction of future states of corrosion defect depth growth
11	The ANN is not suitable to predict very long-term corrosion, but it was expected that its modelling capability could increase with the consideration of more affecting factors and the availability of more accurate data
11	The ANN is not suitable to predict very long-term corrosion, but it was expected that its modelling capability could increase with the consideration of more affecting factors and the availability of more accurate data
12	BPNN was applied to lab data to establish a phenomenological model correlating influential enviromental factors with the pitting risk (characterized by Ecorr - Epit)
13	The ANN ensemble model outperformed the single ANN for predicting CR
13	The ANN ensemble model outperformed the single ANN for predicting volume loss
14	RF combined high flexibility with less data sensitivity, and was able to highly generalize on the corrosion initiation time datasets with high accuracy
14	The results of the effect of split percentage on the performances of the proposed RF revealed that the model variance reduced with an increase in training dataset
6	The ANN (backpropagation) was able to learn the underlying functions in mapping corrosion variables and parameters to different corrosion metrics
6	The ANN (backpropagation) was able to learn the underlying functions in mapping corrosion variables and parameters to different corrosion metrics
7	ANN was used as a predictive tool to investigate the material susceptibility to localized corrosion as a function of alloy treatment and different environmental conditions (pH, [Cl-])
15	Sensitivity analysis was done to evaluate the contribution of each independent variable to the dependent variable. The CEFM-predicted (deterministic model) and ANN-predicted crack growth rates were virtually identical
16	ML showed negligible correlation between quantum chemically derived parameters and corrosion inhibition, suggesting that published structure-inhibition models in the literature were incorrect (these were presumably based on very small numbers of inhibitors)
16	ML showed negligible correlation between quantum chemically derived parameters and corrosion inhibition, suggesting that published structure-inhibition models in the literature were incorrect (these were presumably based on very small numbers of inhibitors)
16	ML showed negligible correlation between quantum chemically derived parameters and corrosion inhibition, suggesting that published structure-inhibition models in the literature were incorrect (these were presumably based on very small numbers of inhibitors)
16	ML showed negligible correlation between quantum chemically derived parameters and corrosion inhibition, suggesting that published structure-inhibition models in the literature were incorrect (these were presumably based on very small numbers of inhibitors)
17	The results indicated that ANN was not able to fit the corrosion performance with the indicator of pH and TA whereas it could present a relatively accurate prediction based on TA/pH
18	The tiering ensemble models outperformed all individual models. Although neither the best baseline machine learner nor the meta ensembles reached a reasonable MAPE, the proposed SFA-LSSVR model was effective (the lowest MAPE values). The SFA was effective for the dynamic optimization of the LSSVR hyperparameters
18	The tiering ensemble models further improved the predictive accuracy of SVR (best individual model). The SFA-LSSVR model outperformed all the others (the lowest MAPE values). The SFA-LSSVR was effective for the dynamic optimization of the hyperparameters
19	Combination of Kohonen mapping (KSOM) and ANN analysis found the most probable values for missing data and uncovered hidden relationships between CT and 17 variables

Reference	Contributions
1	Identification of atmospheric corrosion key factors: Cr, C, Si, P, S, O2_max, S_max, pH_min, T_mean
1	The nine new principal component features obtained by the PCA analysis have no physical meaning
1	The feature creation method I (16 new features) proposed an effective way to simulate the influence of elemental contents and their interactions on steel corrosion resistance
1	The feature creation method II (32 new features) proposed an effective way to simulate the influence of elemental contents and their interactions on steel corrosion resistance
2	During the whole exposure period, the ELEMENTS feature was among the most significant ones; concerning the environmental features, the CHLORIDE and PRECIPIT features were the most significant ones in the first 3 years, and RH_MIN became the most significant factor after 5 years
3	The environmental factor was more important than the material factor (pH being the most important feature at the 2 corrosion stages (time split at Year 5)). Initial stage: the environment was always more important than 80% (SO2/Cl- higher importance than T/RH; Rainfall was the 2nd most important in Year1). Stationary stage (+5 years): corrosion products shielded the LAS, increasing the material importance (SO2 importance lesser than Cl-/T/RH)
4	T, RH and rainfall status were the most important features. Pollutants were not important factors due to their low concentrations
4	T, RH and rainfall status were the most important features. Pollutants were not important factors due to their low concentrations
5	The parameter eta (RGANGBM(1,1)) can reflect the corrosivity of atmospheric environment
6	The ANN model was developed using data organized according to the differences in the environmental conditions, including temperature, pH, etc (each condition was unique)
7	The observed validity of the Arrhenius rule makes the model applicable in the general sense of thermal properties (i.e. Arrhenius parameters and CRs at temperature values outside the range considered)
8	Using small-sample datasets, the SVR model presented a high accuracy for predicting the CR of 3C steel under different seawater environment
9	Knowledge mining could provide quantitative thresholds for the environmental variables: pH = ~6.3; RH = 67% and 86% (electrolyte continuity and film saturation); SO2 = ~0.067 mg/cm3; rainfall = ~320 mm/month; Cl- = 0.8 and 1.9 mg/cm3
10	Multivariate polynomial regression provided a baseline for generating the dataset for training machine learning model
10	Understanding of the expected time-dependent changes (over a 10 years period) in the corrosion defect depth growth trajectory using operating parameters of the pipelines
11	The corrosion rates of steel was mostly like affected by changes in corrosion product: it slowed at high TOW, suggesting that moisture inhibited corrosion due to the formation of protective carbonate; and significantly increased at high SO2 (from sparingly soluble basic sulphates to soluble sulphates)
11	The corrosion rates of zinc slowed at high TOW (confirming that moisture inhibits zinc corrosion because of the formation of protective carbonate films) and significantly increased at high sulphur dioxide concentrations
12	To enhance the chloride threshold of rebar the following measures should be applied: (1) reduce the initial [Cl]); (2) increase the percent of bound chloride; (3) increase the pH of concrete; and (4) minimize the DO
13	By considering several network inputs (geometrical pipeline features, fluid dynamic multiphase variables and deterministic models), the regions of pipelines where corrosion was the most likely were correctly identified
13	By considering several network inputs (geometrical pipeline features, fluid dynamic multiphase variables and deterministic models), the regions of pipelines where corrosion was the most likely were correctly identified
14	The proposed RF model provided excellent approximation of the corrosion initiation time (determined from corrosion potential measurement) of embedded steel by systematically and efficiently learning the model features
14	The RF ensemble model, amongst other trained models, performed best with 85/15 dataset percentage split
6	The ANN model developed using polarization data for general corrosion of metallic glasses showed good agreement with experimental data as a function of the changing environment
6	The ANN model developed using crevice corrosion data on Ti-2 provided the alloy's penetration depth over much longer times (in comparison to the experimental data available)
7	The input variable with the highest weighting was the chloride concentration, which is very realistic
15	Both ANN and the CEFM model predicted that CGR increased strongly with increasing concentration of O2. The CEFM was demonstrated to accurately capture the electrochemical/mechanical/microstructural character of crack growth (validation based on the high fidelity prediction in comparison to ANN)
16	Robust models were developed linking molecular properties calculated by non-quantum chemical methods and corrosion inhibition. This was the most comprehensive inhibitor modelling study, in terms of N° and diversity of inhibitors, range of alloys and pHs
16	Robust models were developed linking molecular properties calculated by non-quantum chemical methods and corrosion inhibition. This was the most comprehensive inhibitor modelling study, in terms of N° and diversity of inhibitors, range of alloys and pHs
16	Robust models were developed linking molecular properties calculated by non-quantum chemical methods and corrosion inhibition. This was the most comprehensive inhibitor modelling study, in terms of N° and diversity of inhibitors, range of alloys and pHs
16	Robust models were developed linking molecular properties calculated by non-quantum chemical methods and corrosion inhibition. This was the most comprehensive inhibitor modelling study, in terms of N° and diversity of inhibitors, range of alloys and pHs
17	The results indicated that the corrosion performance of the PCC increased with the decrease in TA/pH. However, it was impossible to produce a coating with good corrosion performance in pure water (TA/pH equal to 0). There is a critical value of TA/pH which should be the boundary of the theory of TA/pH
18	The proposed SFA-LSSVR model was effective in modeling pitting risk of rebar steel
18	The proposed SFA-LSSVR model was effective in modeling corrosion rate of carbon steels
19	ANN found that CT is only poorly correlated with porosity, w/c ratio, and the Al, Fe, K, Mg, and S oxide contents of the cement

Reference	Critical points
1	Features such as the flow velocity or microorganisms, also seriously influencing the corrosion rate of LAS in marine environments, were not explored
1	Features such as the flow velocity or microorganisms, also seriously influencing the corrosion rate of LAS in marine environments, were not explored
1	Modelling including the top 4 most important physical features (creation method I) produced lower performance than considering the top 5 elements instead
1	Modelling including the top 4 most important physical features (creation method II) produced lower performance than considering the top 5 elements instead
2	Low generalization ability: the data amount was insufficient to fully reflect the impact of all factors; data bias: low accuracy for samples with high CRs (necessity of adding more data with high CRs)
3	Need for features describing the material microstructure
4	The sensor ended almost fully covered with corrosion products, which affected its response over time
4	The deposition of chlorides (monthly recording) was not considered as a feature
5	The results from BPNN and SVR (16th year) were unreliable due to the small sample dataset (despite training with LOOCV)
6	The training-testing split method is not clear. The testing dataset seems to include only one type of steel. The training performance is not presented
7	Results presented considered the feature exposure time as a constant only. The same data was used for training-testing (the employed CV method was not specified)
8	The SVR model presented a limited extrapolation ability. The collection of more experimental data would enhance its regression accuracy. Unknown electrochemical methods for Data acquisition
9	The paper did not discuss the effects of composition/microstructure. The models are not directly comparable with those presented in the preliminary work because 1 more input was here considered (Fe)
10	The multivariate polynomial regression good performance was most likely linked to the large input dataset (8300 samples)
10	The final data dimension upon application of PCA is not presented
11	The high variances obtained with ANN were mainly due to two reasons: ignorance of other important affecting variables and inherent scatter in the data from a number of sources
11	The high variances obtained with ANN were mainly due to two reasons: ignorance of other important affecting variables and inherent scatter in the data from a number of sources
12	Knowledge mining: only the overall trend with independent variables maintained at specific levels were discussed. The influence of these variables and their interactions on the pitting risk of steel in simulated concrete pore solutions is very complex
13	The small size of the training sample for the network modelling. Predictions could be further improved by considering larger datasets (several pipeline cases and flow conditions)
13	The small size of the training sample for the network modelling. Predictions could be further improved by considering larger datasets (several pipeline cases and flow conditions)
14	Tests were performed under laboratory conditions only. It would be recommended to validate the ensemble ML techniques applied to the prediction of concrete corrosion from in-field samples
14	Tests were performed under laboratory conditions only. It would be recommended to validate the RF ensemble method applied to the prediction of concrete corrosion from in-field samples
6	Train-test split procedure not clear
6	Due to the small number of experimental data points available, the data was not divided into training and test sets (only 8 samples were used for training/validation)
7	Only the average R^2 value (from CV) was provided: no distinction between training and testing performances. Small sample dataset
15	The flow velocity was not included in the analysis, even though theory indicates that it is an important variable
16	The number of descriptors necessary to generate useful models, and the relatively opaque or arcane nature of many of them makes mechanistic interpretation of the models problematic
16	The number of descriptors necessary to generate useful models, and the relatively opaque or arcane nature of many of them makes mechanistic interpretation of the models problematic
16	The number of descriptors necessary to generate useful models, and the relatively opaque or arcane nature of many of them makes mechanistic interpretation of the models problematic
16	The number of descriptors necessary to generate useful models, and the relatively opaque or arcane nature of many of them makes mechanistic interpretation of the models problematic
17	It is very complex to decrease TA/pH under the premise of maintenance stability of the conversion bath. The validity of using TA/pH in large pH range should be discussed in details
18	The hybrid model obtained a relatively high MAPE value for pitting risk (5.60%) in comparison to corrosion rate (1.26%). Further study is needed to investigate the prediction accuracy and model applicability of the SFA-LSSVR
18	Further study is needed to investigate the prediction accuracy and model applicability of the SFA-LSSVR. Dataset derived from literature not in English. Unknown electrochemical methods for Data acquisition
19	The correlations between dependent and independent variables depend on which definition of CT is employed, because of concomitant changes in pH

Supplementary References

1. 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).

2. 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. Mater.* **21**, 359–370 (2020).

3. Zhi, Y., Fu, D., Zhang, D., Yang, T. & Li, X. Prediction and Knowledge Mining of Outdoor Atmospheric Corrosion Rates of Low Alloy Steels Based on the Random Forests Approach. *Metals (Basel).* **9**, 383 (2019).

4. 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).

5. Zhi, Y. *et al.* Long-term prediction on atmospheric corrosion data series of carbon steel in China based on NGBM(1,1) model and genetic algorithm. *Anti-Corrosion Methods Mater.* **66**, 403–411 (2017).

6. Kamrunnahar, M. & Urquidi-Macdonald, M. Prediction of corrosion behavior using neural network as a data mining tool. *Corros. Sci.* **52**, 669–677 (2010).

7. Kamrunnahar, M. & Urquidi-Macdonald, M. Prediction of corrosion behaviour of Alloy 22 using neural network as a data mining tool. *Corros. Sci.* **53**, 961–967 (2011).

8. Wen, Y. F. *et al.* Corrosion rate prediction of 3C steel under different seawater environment by using support vector regression. *Corros. Sci.* **51**, 349–355 (2009).

9. Zhi, Y., Yang, T. & Fu, D. An improved deep forest model for forecast the outdoor atmospheric corrosion rate of low-alloy steels. *J. Mater. Sci. Technol.* **49**, 202–210 (2020).

10. Ossai, C. I. A Data-Driven Machine Learning Approach for Corrosion Risk Assessment—A Comparative Study. *Big Data Cogn. Comput.* **3**, 28 (2019).

11. Cai, J., Cottis, R. A. & Lyon, S. B. Phenomenological modelling of atmospheric corrosion using an

artificial neural network. *Corros. Sci.* **41**, 2001–2030 (1999).

12. Shi, X., Anh Nguyen, T., Kumar, P. & Liu, Y. A phenomenological model for the chloride threshold of pitting corrosion of steel in simulated concrete pore solutions. *Anti-Corrosion Methods Mater.* **58**, 179–189 (2011).

13. De Masi, G., Gentile, M., Vichi, R., Bruschi, R. & Gabetta, G. Machine learning approach to corrosion assessment in subsea pipelines. in *OCEANS 2015 - Genova* 1–6 (IEEE, 2015). doi:10.1109/OCEANS-Genova.2015.7271592.

14. 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).

15. Shi, J., Wang, J. & Macdonald, D. D. Prediction of crack growth rate in Type 304 stainless steel using artificial neural networks and the coupled environment fracture model. *Corros. Sci.* **89**, 69–80 (2014).

16. Winkler, D. A. *et al.* Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corros. Sci.* **106**, 229–235 (2016).

17. Chunyan, Z. *et al.* Ratio of total acidity to pH value of coating bath: A new strategy towards phosphate conversion coatings with optimized corrosion resistance for magnesium alloys. *Corros. Sci.* **150**, 279–295 (2019).

18. Chou, J., Ngo, N. & Chong, W. K. The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate. *Eng. Appl. Artif. Intell.* **65**, 471–483 (2017).

19. Zhu, Y., Macdonald, D. D., Qiu, J. & Urquidi-Macdonald, M. Corrosion of rebar in concrete. Part III: Artificial Neural Network analysis of chloride threshold data. *Corros. Sci.* **185**, 109438 (2021).