Supplementary Information (SI) for Environmental Science: Advances.

Machine learning-based assessment of arsenic-contaminated groundwater quality in the Red River Delta: a case study in Van Phuc, Hanoi, Vietnam

Vu Thi Duyen,[a,b] Nguyen Thanh Dam,[c] Pham Thi Kim Trang,[d] Michael Berg[e] and Pham Hung Viet[a,*]

a. Key laboratory of Analytical Technology for Environmental Quality and Food Safety Control, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan District, Hanoi 100000, Vietnam.

b. Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay District, Hanoi 100000, Vietnam.

c. Faculty of Chemistry, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan District, Hanoi 100000, Vietnam.

d. Research Center for Environmental Technology and Sustainable Development, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan District, Hanoi 100000, Vietnam

e. Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland

Email: vietph@vnu.edu.vn

Supplementary Information

- Tables S-1 to S-6

Supplementary Tables

Table S-1. Descriptive statistics of 11 selected groundwater quality parameters in Van Phuc together with the Vietnamese national guideline values stipulated in the technical regulation on groundwater quality.

Parameters	Unit	Min	Max	Mean	SD	Guideline values	
						Lower	Upper
pH	-	6.60	8.40	7.28	0.36	5.8	8.5
As	mg/L	0	0.53	0.17	0.19	-	0.05
Fe	mg/L	0	24.83	9.40	8.18	-	5.0
Mn^{2+}	mg/L	0	3.38	1.13	0.94	-	0.5
Na^{+}	mg/L	2.67	43.41	17.63	10.45	-	200
NH_4^{+}	mg-N/L	0	67.40	13.88	15.84	-	1.0
F^{-}	mg/L	0.06	2.36	0.35	0.44	-	1.5
Cl^{-}	mg/L	4.40	64.00	22.49	15.23	-	250
SO_4^{2-}	mg/L	0	29.90	1.73	5.61	-	400
TDS	mg/L	278.05	949.39	628.02	156.17	-	1,500
HN	mg/L	113.51	576.13	362.57	13.88	-	500

Table S-2. Groundwater dataset used for machine learning modelling: concentrations of 11 physico-chemical groundwater parameters and computed pollution status. Highlighted cells mark samples exceeding guideline thresholds.

Sample	Depth	pH	As	Fe	Mn^{2+}	Na^{+}	NH_4^{+}	F^{-}	Cl^{-}	SO_4^{2-}	TDS	HN	PS
	m.g.b.l		mg/L	mg/L	mg/L	mg/L	mg N/L	mg/L	mg/L	mg/L	mg/L	mg/L	
VP1	-25	7.74	0.14	12.58	0.77	6.13	0.60	0.07	5.55	29.90	623.10	433.61	1
VP2	-20	6.65	0.08	17.89	0.34	39.65	10.65	0.08	64.00	0.00	928.62	564.22	1
VP3	-30	7.45	0.24	22.14	0.23	14.14	15.60	0.08	24.58	0.00	694.79	455.38	1
VP4	-40	7.46	0.24	23.54	0.29	21.11	12.34	0.12	45.03	0.00	807.35	485.85	1
VP5	-20	7.26	0.44	9.97	1.05	9.29	53.00	0.15	11.08	0.00	697.47	297.75	1
VP6	-28	7.31	0.52	12.16	0.27	14.01	33.80	0.14	11.29	0.00	639.18	293.77	1
VP7	-40	7.36	0.39	13.60	0.00	12.39	29.73	0.16	13.45	0.00	670.67	345.77	1
VP8	-27	7.25	0.34	20.20	0.00	16.38	5.69	0.06	39.48	0.00	690.77	453.21	1
VP9	-20	7.12	0.44	20.09	0.00	12.62	5.24	0.08	32.06	0.00	708.86	468.19	1
VP10	-25	7.44	0.53	12.22	0.00	10.90	67.40	0.13	5.07	0.00	840.18	348.13	1
VP11	-37	7.36	0.42	11.40	0.21	16.24	10.17	0.11	20.99	0.00	703.50	440.49	1
VP12	-25	7.19	0.42	11.46	0.55	11.69	26.50	0.25	12.44	0.00	680.72	386.16	1
VP13	-32	7.38	0.00	0.00	1.59	15.14	9.46	0.42	28.21	0.00	665.98	425.44	1
VP14	-47	6.60	0.01	16.00	1.14	43.41	0.54	0.16	14.15	0.57	347.73	133.75	1
VP15	-25	7.36	0.29	10.44	0.96	10.88	18.20	0.34	24.76	0.00	580.89	338.13	1
VP16	-29	7.32	0.05	9.45	2.36	10.16	14.60	0.47	19.72	0.00	585.58	368.36	1
VP17	-37	7.11	0.00	0.00	2.69	17.84	0.97	0.39	18.82	0.00	629.13	429.63	1
VP18	-32	7.18	0.00	0.68	2.06	12.83	5.71	0.34	34.33	0.47	653.25	441.84	1
VP19	-42	6.99	0.01	1.85	2.24	18.13	1.71	0.29	15.17	0.00	554.09	400.33	1
VP20	-25	7.83	0.07	8.46	3.38	9.49	15.30	0.46	15.95	0.00	587.59	353.86	1
VP21	-25	7.14	0.00	0.73	1.99	11.38	11.04	2.36	30.33	0.00	616.40	394.71	1
VP22	-24	7.24	0.00	0.00	1.26	10.63	11.93	0.85	24.47	0.00	577.54	379.31	1
VP23	-38	7.09	0.00	0.00	1.65	13.73	3.17	0.36	30.15	0.00	635.83	482.02	1
VP24	-22	6.93	0.00	0.03	2.49	30.59	0.00	0.34	4.51	4.74	347.73	177.97	1
VP25	-38	6.95	0.00	0.00	0.32	32.70	0.00	0.26	4.40	6.28	278.05	122.35	0
VP26	-54	7.08	0.01	24.83	1.14	27.65	0.65	0.33	9.17	1.47	404.01	187.82	1
VP27	-25	7.82	0.04	1.40	0.00	2.67	10.25	0.15	16.42	0.00	671.34	113.16	1
VP28	-25	8.40	0.02	0.56	1.61	19.25	23.65	0.97	16.01	0.44	442.87	166.92	1
VP29	-48	7.17	0.08	10.49	1.46	40.27	4.51	0.13	60.6	2.97	949.39	573.95	1
S1	-40	6.88	0.03	0.86	0.22	3.29	0.72	0.63	2.15	7.95	117.32	73.28	0
S2	-29	6.80	0.01	3.74	0.39	9.42	0.00	0.78	3.23	7.21	118.99	66.37	0
S3	-40	7.50	0.00	0.91	0.00	2.57	0.32	0.19	2.68	8.91	136.68	87.74	0
S4	-17	6.69	0.00	0.00	0.00	22.96	0.00	0.08	29.91	65.28	568.83	452.88	0
S5	-40	6.18	0.00	4.11	0.26	40.71	0.00	0.11	14.69	0.86	147.40	19.89	0
S6	-20	6.40	0.00	0.51	0.28	58.95	0.00	0.17	30.57	8.54	211.05	17.31	0
S7	-45	6.28	0.00	0.00	0.00	24.01	0.23	0.13	1.99	0.47	115.91	28.43	0

Table S-3. Simple hyperparameter grids used for tuning decision tree, random forest, and XGBoost algorithms

Algorithm	Hyperparameter	Value
Decision tree	cp	c(0.001, 0.1, 10)
Random forest	mtry	c(1, 10, 10)
	ntree	500
XGBoost	nrounds	c(100, 200)
	max_depth	c(3, 5)
	learning rate	c(0.01, 0.05)
	colsample_bytree	c(0.5, 0.8)
	subsample	c(0.5, 0.8)
	gamma	0
	min_child_weight	1

Table S-4. Model comparison metrics for decision tree, random forest, and XGBoost algorithms

Model	Dataset	AUC	Logloss	Accuracy	Sensitivity	Specificity
Decision Tree	Train	0.500	0.693	0.214	0	1
Decision Tree	Test	0.500	0.693	0.250	0	1
Random Forest	Train	0.992	0.207	0.929	0.901	1
Random Forest	Test	1	0.299	0.875	0.833	1
XGBoost	Train	1	0.359	0.964	0.955	1
XGBoost	Test	1	0.351	1	1	1

Table S-5. Hyperparameter grid used for tuning the XGBoost classifier.

Model hyperparameter	Grid search value									
learning rate	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
max_depth	1	2	3	4	5	6	7	8	9	10
subsample	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
colsample_bytree	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
scale_pos_weight	0.25	1								
min_child_weight	1									
gamma	0									

Table S-6. Five-fold cross-validation performance of the optimised XGBoost model.

Fold	Train AUC	Test AUC	Train logloss	Test logloss	Best iterration
1	1	0.875	0.235	0.477	185
2	1	1	0.196	0.283	243
3	1	1	0.204	0.178	495
4	1	1	0.204	0.226	256
5	1	1	0.216	0.165	346