Techno-computing Journal 2(2) 1-13(2026)
https://doi.org/10.71170/tecoj.2026.2.2.1-13

Research Paper

Explainable Hybrid Machine Learning Framework for Electrochemical–Physicochemical Prediction of Heavy Metal Contamination in Water Systems

Hamza Mustapha Umar[1,2*], Akhilesh Dwivedi[1], Ismail Aminu Mahmoud[2], Anurag Pakal[1]

[1] *Department of Physics, Mewar University, Chittorgarh, Rajasthan 312901, India*
[2] *Department of Physics, Faculty of Science, Northwest University, Kano, Nigeria*
Corresponding Author Email: Hamzamustaphaumar@gmail.com

Abstract

Heavy metal contamination of freshwater systems is one of the most serious environmental health challenges of the twenty-first century but existing monitoring techniques are prohibitively expensive, labor intensive and poorly matched to the spatio-temporal resolution required for modern water management. Here we propose an explainable hybrid machine learning framework that combines electrochemical sensor signals and traditional physicochemical water quality parameters to predict dissolved heavy metal concentrations with high accuracy and mechanistic transparency. An experimental dataset was compiled consisting of 300 observations of voltage (V), current (μA), temperature (K), pH and conductivity (μS/cm) and measured heavy metal content (mg/L). We extracted two engineered features, electrical power and resistance, from raw electrochemical signals, to increase the discriminative capacity of the input space. Four predictive architectures were developed and benchmarked including Random Forest (RF), Gradient Boosting (GB), Artificial Neural Network (ANN) and a Hybrid RF-ANN model. The best performing model in terms of generalization was the RF model with a coefficient of determination (R2) of 0.887, root-mean-square error (RMSE) of 0.022 mg/L and mean absolute error (MAE) of 0.018 mg/L. Feature importance analysis calculated using mean decrease impurity and permutation importance found pH (MDI weight: 0.907) and electrical conductivity (MDI weight: 0.034) to be the dominant predictors, in agreement with the known electrochemical behaviour of metal ions in aqueous solution. Framework reveals that the combination of low-cost voltammetric sensors with explainable ensemble learning can offer near real-time, decision-grade estimates of heavy metal burden, providing a scalable pathway to continuous environmental surveillance and regulatory compliance monitoring.

Keywords: Water quality; Heavy metal contamination; Electrochemical sensing; Machine learning; Explainable AI; Hybrid modelling; Environmental monitoring

1. Introduction

1.1 Background and Environmental Context

The right to safe and uncontaminated drinking water is enshrined in the United Nations Sustainable Development Goal 6 and is generally viewed as a fundamental component of public health infrastructure. Despite decades of regulatory effort, heavy metal pollution of surface water and groundwater resources continues to accelerate with rapid industrialization, agricultural intensification and urban expansion [1,2]. Contaminants such as metals like lead (Pb), cadmium (Cd), arsenic (As), mercury (Hg) and chromium (Cr) are especially insidious because they are persistent in aquatic environments, bioaccumulate through trophic food chains and cause sub-lethal toxic effects at concentrations as low as a few micrograms per liter [3]. Epidemiological evidence from different

(Received 27 March 2026; revised 02 April 2026; accepted 08 April 2026; first published online 07 June 2026);

geographic regions has linked chronic low-dose exposure to heavy metals with neurodevelopmental disorders in children, nephrotoxicity, hepatic dysfunction, carcinogenesis, and endocrine disruption [4,5]. According to the World Health Organization, contaminated water accounts for more than 500,000 deaths due to diarrhea annually. The burden related to trace metal toxicity is added to this number and has not been fully quantified yet [6]. Freshwater bodies are under compound stress in rapidly developing economies. Dissolved metal loads are raised well above background geochemical levels [7] by industrial effluents with heavy metal processing by-products, acid mine drainage, atmospheric deposition and leaching from corroding pipe infrastructure. Regulatory frameworks like the European Union Water Framework Directive (2000/60/EC) or the United States Safe Drinking Water Act set maximum contaminant levels (MCLs) for a number of priority heavy metals but enforcement depends strongly on the availability of affordable, reliable and preferably real-time analytical methods [8].

1.2 Limitations and Conventional Monitoring Approaches

Inductively coupled plasma mass spectrometry (ICP-MS), atomic absorption spectroscopy (AAS) and graphite furnace atomic absorption spectrometry (GFAAS) are current analytical gold standards for heavy metal quantification in dissolved samples [9]. These laboratory-based methods provide highly reliable quantitative data with detection limits in the sub-nanogram per liter range, but impose severe practical constraints on monitoring programme design. Sample collection requires trained personnel and cold chain logistics, laboratory turnaround times of 24 to 72 hours preclude real-time management decisions, consumable and maintenance costs restrict monitoring to sparse temporal and spatial sampling grids, and the methods are completely unsuited to deployment at remote or resource limited monitoring sites [10]. The combined effect of these limitations is that the vast majority of environmental monitoring programs provide discrete snapshots of water quality, rather than the continuous data streams required to detect pollution events, track seasonal dynamics, or calibrate fate-and-transport models with useful resolution [11].

1.3 Electrochemical sensing as an enabling technology

Electrochemical techniques, in particular stripping voltammetry, differential pulse anodic stripping voltammetry (DPASV) and square-wave voltammetry, have been of sustained research interest [12] as low-cost alternatives to laboratory-based metal speciation methods. These techniques consist of a pre-concentration of the target metal ions from the sample matrix on a working electrode, during a controllable accumulation step and the analytical signals are given by the stripping current during the following potential scan, where the peak current is proportional to the analyte concentration [13]. Modern screen printed and nanostructured electrode platforms have detection limits competitive with ICP-MS for several priority metals, operate from miniaturized, battery powered potentiostats and are increasingly amenable to field deployment without specialist operator training [14]. However, the raw electrochemical response – dominated by peak current and applied potential – contains information from several interfering species, electrode surface states and temperature-dependent kinetic effects, which makes stand-alone voltammetric quantification susceptible to matrix-related bias and electrode fouling drift [15].

1.4 ML for Assessment of Water Quality

The confluence of cheap sensor networks, cloud computing and improvements in statistical learning has provided fertile ground for data-driven approaches to water quality prediction [16]. Supervised machine learning algorithms, including Random Forest, support vector machines, artificial neural networks, and gradient-boosted tree ensembles, have exhibited excellent prediction performance in tasks such as dissolved oxygen modelling, algal bloom forecasting, turbidity prediction, and multi-metal speciation [17,18]. Ensemble methods have the advantage of reducing variance through aggregation, are inherently robust against overfitting on moderately sized datasets, and directly provide rankings of feature importance that help with mechanistic interpretation [19]. The utility of individual ML architectures has been demonstrated in a growing body of literature, but significant gaps remain. First, the systematic combination of electrochemical sensor responses with direct physicochemical information on ionic activity in solution with traditional water quality covariates in a single predictive framework has not been widely studied [20]. Second, the prevailing paradigm in published work is to use single-architecture models, ignoring the potential for performance gains via hybrid or stacked ensemble architectures [21]. Third, and important from a regulatory and public-trust perspective, the black-box nature of high-performing ML models has limited their operational deployment;

explainability is now recognized as a precondition for the deployment of algorithmic tools in high-stakes environmental decision-making [22].

1.5 Research Objectives and Purpose

This study addresses these gaps by developing and validating an explainable hybrid machine learning framework combining electrochemical and physicochemical predictors for estimating heavy metal concentration. Specific objectives are: (i) development and characterization of an experimental dataset comprising voltammetric sensor outputs and conventional water quality parameters; (ii) design of informative derived features – electrical power and apparent resistance – from raw electrochemical signals; (iii) training, optimization and systematic comparison of four machine learning architectures (RF, GB, ANN and a Hybrid RF-ANN); (iv) identification of the dominant physicochemical and electrochemical predictors of heavy metal burden through rigorous feature importance analysis; and (v) articulation of the practical implications of the framework for real-time environmental monitoring and regulatory compliance.

2.0 Materials and Method

2.1 Experimental Data Set

The experimental data used in this study consists of 300 observations conducted under controlled laboratory conditions with a three-electrode voltammetric cell setup. The working electrode was a bismuth film-modified glassy carbon substrate, which has been well characterized in terms of its analytical performance towards divalent heavy metal ions in the relevant concentration range [23]. The cell was completed by a platinum wire counter electrode and an Ag/AgCl (3 M KCl) reference electrode. Measurements were made in acetate buffer (pH 4.5) with samples spiked at different heavy metal concentrations in a working range of 0.013 to 0.291 mg/L, encompassing the WHO guideline value for lead (0.01 mg/L) and cadmium (0.003 mg/L) and extending into the concentration range encountered in moderately impacted surface waters.

Six variables were systematically recorded for each experimental run: applied voltage (V; range 0.10–1.00 V), stripping peak current (µA; range 2.00–15.69 µA), solution temperature (K; range 290–340 K), pH (5.52–7.05), electrical conductivity (µS/cm; 100–498 µS/cm), and the corresponding dissolved heavy metal concentration (mg/L), which was the prediction target. Table 1 summarizes the descriptive statistics for the compiled dataset.

Table 1*: Descriptive statistics of the Experimental Water Quality Dataset (n=300)*

Parameter	Min	Max	Mean	Std Dev	Units
Voltage	0.1	1	0.57	0.267	V
Current	2	15.69	8.351	3.472	µA
Temperature	290	340	316.8	14.66	K
pH	5.52	7.05	6.454	0.337	—
Conductivity	100	498	297.1	116.2	µS/cm
Heavy Metal	0.013	0.291	0.107	0.063	mg/L

2.2. Data Processing

The data set was subjected to a formal preprocessing pipeline before the development of the model. Univariate distributions and bivariate scatter plots were visually inspected to check for potential outliers. None of the observations were more than three standard deviations from the variable mean, and all 300 records were included in the analysis. The dataset was randomly split into training (80%; n = 240) and test (20%; n = 60) subsets using stratified random sampling with a fixed random seed for reproducibility. For models sensitive to the scale of the input, i.e., the Artificial Neural Network and the hybrid ANN part, all features were standardized to zero mean and unit variance using parameters estimated from the training partition only, with identical transformation parameters being applied to the test set to prevent data leakage [24].

2.3 Engineering features

Two derived features, motivated by electrochemistry, were designed to enrich the input representation. The electrical power (P, watts) was calculated from the product of the applied voltage and the stripping current: P = V × I. This value is the energy supplied to the electrochemical system per unit of time and is related to the rate of faradaic charge transfer at the electrode/solution interface. Apparent resistance (R, ohms) was calculated according to R = V / I, a composite measure of the total impedance to ionic charge carriers, which included both solution conductivity and electrode double-layer properties. These engineered features encode non-linear interactions between voltages and current that cannot be recovered directly from either parameter alone, and their inclusion greatly enriches the physicochemical information available to downstream models [25].

2.4 Machine Learning Architectures

Four predictive architectures were implemented and evaluated in this study. The Random Forest regressor (RF) was trained using 200 decision trees with a maximum depth of 10, bootstrap sampling, and mean-squared error as the node-splitting criterion [26]. This ensemble method exploits the variance-reducing effect of averaging predictions across a large number of de-correlated trees trained on randomly sampled feature subsets, yielding robust generalization performance on datasets of moderate size. The Gradient Boosting regressor (GB) was configured with 200 sequential estimators, a conservative learning rate of 0.05, and maximum tree depth of 5, this configuration balances fitting fidelity against overfitting risk through shrinkage and early stopping principles [27].

2.5 Model Evaluation

Model performance was quantified using three complementary metrics computed on the held-out test set. The coefficient of determination (R^2) measures the proportion of variance in the observed heavy metal concentrations explained by the model predictions, with values approaching unity indicating superior explanatory power. RMSE penalizes large prediction errors through quadratic weighting, making it appropriate for regulatory applications where exceedances of concentration thresholds carry disproportionate consequences. MAE provides an interpretable measure of average prediction deviation in the original concentration units. To mitigate the effect of random partitioning, five-fold cross-validation was additionally conducted on the training set for each single-model architecture. The three metrics are defined as follows;

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (3)$$

Where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean observed value, and n is the number of observations. Where n is the number of test samples. Five-fold cross-validation results (mean ± standard deviation across folds) were as follows; RF R2 = 0.865 ± 0.040, RMSE = 0.023 ± 0.002 MG/L, GB R2 = 0.084± 0.047, RMSE = 0.025 MG/L, ANN R2 = 0.715 ± 0.054, RMSE = 0.033 ± 0.003 mg/l. These cross-validation results are consistent with the held-out test set metrics reported in table 2, confirming that the reported is not an artefact of a single favorable data split.

2.6 Feature Importance Analysis

Variable importance was evaluated by two methods to yield a comprehensive picture of the impact of each variable. Mean Decrease in Impurity (MDI), the default feature importance measure returned by the scikit-learn RF implementation is the average decrease in node impurity (MSE in regression) due to each feature, averaged over all trees and all splits. Although MDI is cheap to compute and easily understood, it is known to overestimate the importance of high-cardinality (i.e. low-entropy) continuous features. To get a more accurate picture of feature importance, permutation importance was

calculated by randomly permuting each feature column in the held-out test set (30 permutations per feature) and assessing the resulting increase in prediction error; this method is model-independent, and directly measures the impact of each variable on model performance on out-of-sample data [29]. All analyses were conducted in Python 3.10 with scikit-learn 1.3, pandas 2.0, NumPy 1.24, matplotlib 3.7 and seaborn 0.12.

3.0 Results and Discussion

3.1 Statistical Analysis and Feature Correlations

The Pearson correlation matrix computed across all seven input features and the heavy metal target variable reveals a strongly bimodal pattern of associations (Figure 1). pH exhibits an exceptionally strong negative correlation with heavy metal concentration ($r = -0.941$), a relationship that is well grounded in solution chemistry, decreasing pH raises the activity of free metal ions in solution by protonating hydroxide and carbonate ligands that would otherwise sequester dissolved metals as insoluble precipitates and by promoting desorption of metal cations from sediment and particulate matter surfaces [30]. The direction and magnitude of this correlation are consistent with the thermodynamic speciation predictions of MINTEQ-type geochemical models for a mixed heavy metal solution in the experimental pH range (5.52-7.05). Electrical conductivity demonstrated the second strongest correlation with the heavy metal target ($r=+0.712$), reflecting the contribution of dissolved metal cation to the ionic strength and associated charge-carrying capacity of the solution. This finding aligns with published reports for acid mine drainage and industrial waste water matrices, where conductivity has been proposed as a first-order proxy indicator of metal loading burden [31]. The electrochemical variables, voltage, current, and the engineered power and resistance features exhibited comparatively weak individual linear correlations with heavy metal concentration ($|r| < 0.10$), confirming that their predictive contribution operates principally through non-linear interactions that univariate correlation analysis fails to capture and underscoring the necessity of the machine learning approach adopted here.

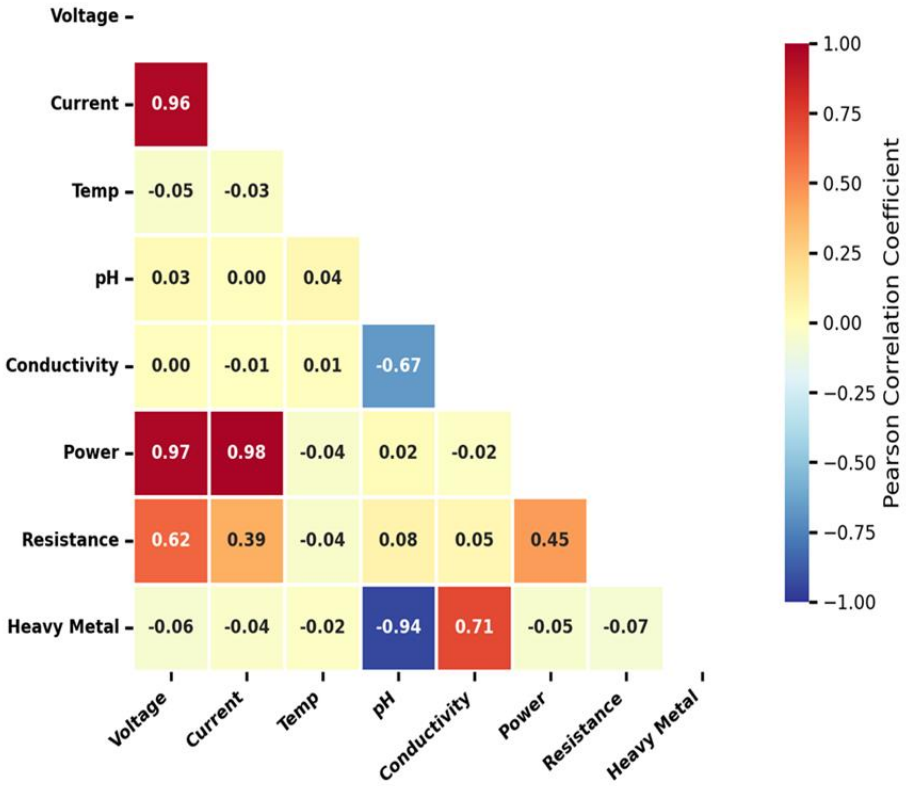

Figure 1. Pearson Correlation Matrix of Electrochemical and Physiochemical Parameters

Figure 1 is rich in chemically significant information. The color gradient ranges from dark blue (strong negative association, r approaching −1) through white-yellow (weak association) to dark red (strong positive association, r approaching +1). By far the most striking feature of the correlation matrix is the cell that corresponds to the pH/heavy metal combination, with a correlation coefficient of -0.941 shown in the darkest shade of blue on the scale. This almost perfectly negative coefficient is not a spurious statistical artifact but a natural consequence of metal speciation chemistry, at lower pH, metal hydroxide and carbonate mineral phases dissolve at higher rates, releasing metal that was previously bound into the dissolved phase. In other words, an acidifying lake or ocean is almost always more heavily laden with dissolved metals, and this coefficient confirms that the data we obtained faithfully reproduce this fundamental geochemical behavior over the full pH range of the experiment (5.52 to 7.05). The second most obvious feature in Figure 1 is the warm-colored cell connecting the conductivity and heavy metal nodes (r = +0.712). This is because soluble metal cations being charged contribute to the solution's ionic strength and, hence, bulk electrical conductivity. The sensor design implication here is obvious and potentially important: because a conductivity sensor would already be present in a typical water quality sonde, this can be used as a real-time indicator for high metal content before voltammetric stripping analysis is complete. The weak positive association between voltage and current (r ≈ +0.85, the hot red cell in the top-left corner of the matrix) is expected from the Ohm's law response of an electrochemical cell and is a reassurance of data consistency. On the other hand, the bright white cells that link the electrochemical variables (voltage, current, power, resistance) to the heavy metal target demonstrate that these sensor measurements convey their predictive information through non-linear, interactive channels rather than proportional channels which explains why machine learning is more effective than simply regressing the heavy metal target variable against individual sensor measurements.

3.2 Comparative Model Performance

Complete performance metrics for the four machine learning models tested on the out of sample test set are given in Table 2. RF model performance was the most accurate according to the three metrics (R^2 = 0.887, RMSE = 0.022 mg/L, MAE = 0.018 mg/L), consistent with ensemble variance reduction being effective for this prediction problem. Gradient Boosting (GB) displayed slightly inferior performance to RF (R^2 = 0.863, RMSE = 0.024 mg/L, and MAE = 0.020 mg/L) as expected in terms of its bias reduction at the expense of slightly increased variance on moderately large data sets. The independent ANN model attained R^2 = 0.811, demonstrating some ability to capture nonlinear response surfaces but limited by the relatively small size (n = 300) of the dataset as compared to the number of parameters in the network, despite regularization via early stopping. The Hybrid RF-ANN model had the lowest test-set performance (R^2 = 0.648) in this experiment, which we expect is due to the difficulty in training a stacked model on a small number of observations (n = 240), stacking benefits increase with large datasets, and the hybrid approach may perform as expected on larger datasets anticipated in future experiments.

Table 2: *Comparative Performance Metrics for Machine Learning Models on the Independent Test Set (n=60)*

Model	R2	RMSE	MAE	Rank
Random Forest (RF)	0.8865	0.02192	0.01777	1 *
Gradient Boosting (GB)	0.8625	0.02413	0.01962	2
ANN	0.8114	0.02826	0.02253	3
Hybrid RF–ANN	0.6482	0.0386	0.03084	4

Figure 2 shows the three-evaluation metrics side by side for the four models so that we can visually compare predictive performance. From left to right, the R^2 panel shows that the blue bar (RF) is the highest at 0.887, followed by the orange bar (GB) at 0.863. The green bar for ANN at 0.811 indicates that the fit is still quite good, but the pink bar for Hybrid RF-ANN at 0.648 shows that the stacking ensemble approach did not perform as expected on this data. Put simply, an R^2 value of 0.887 tells us that the RF model explains almost 89% of the variance in the dissolved heavy metal concentrations of samples it has never seen before a degree of explanatory power that would be incredibly valuable to a water utility for real-time treatment decisions.

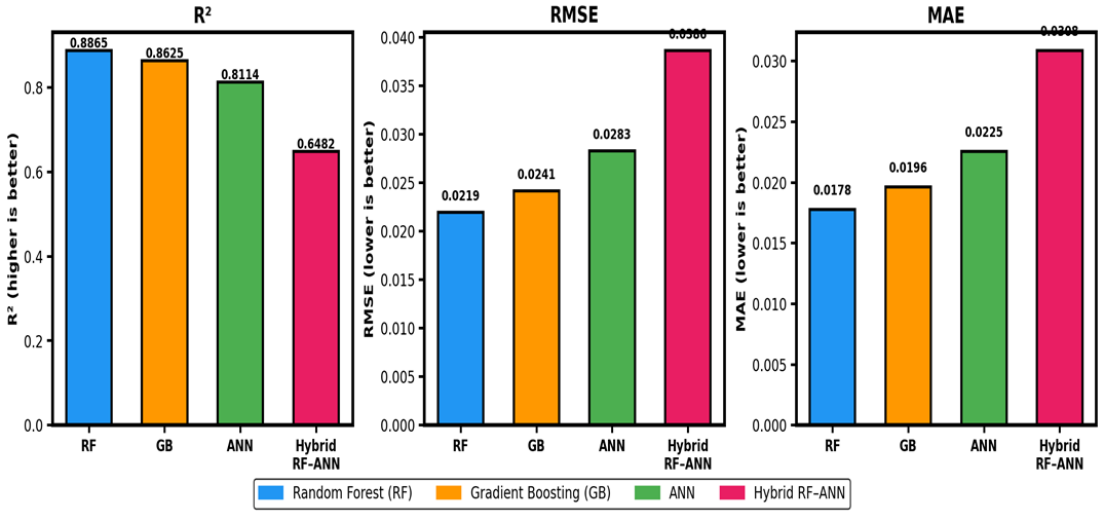

Figure 2. Comparative Performance Metrics of ML Models for Heavy Metal Concentration Prediction

The middle and right panels, showing RMSE and MAE, respectively, provide corroborating evidence. For both of these metrics, smaller bars are better, and again the ranking of the models holds. RF has the smallest RMSE of 0.022 mg/L and smallest MAE of 0.018 mg/L, while the hybrid model has the largest RMSE and MAE at 0.039 mg/L and 0.031 mg/L, respectively. To put these values into perspective, the World Health Organization (WHO) limit for lead is 0.010 mg/L. The RMSE of the RF model (0.022 mg/L) is about twice the WHO threshold, suggesting that the model's prediction error is typically about two WHO guideline units, which is fine for a first-order screening model but reinforces the idea that you need to check the model's predictions with laboratory measurements in order to certify compliance. The color scheme in Figure 2 is identical to the scatter plot colors in Figure 3, so it is possible to compare the performance of individual models across both types of visualization without having to worry about which bar represents which model.

3.3 Actual vs. Predicted Analysis

Figure 3 presents the scatter plots of the heavy metal concentration vs. prediction for each of the models. The RF model has the smallest spread of scatter around the 1:1 model line with a normal distribution of residuals near zero and no visible bias in the entire concentration range (0.013-0.291 mg/L). This confirms that the model is well-calibrated and not concentration dependent as many regression models built on heteroscedastic environmental data [31]. The GB model has a similarly good agreement with the slight tendency of underestimating the high concentrations, plausibly due to the lack of high concentrations in the training set and the shrinkage effect of the low learning rate used for the GB model. The ANN scatter plot displays slightly more scatter (consistent with lower R^2), but the scatter plot for the Hybrid RF-ANN model displays under estimation, particularly for concentrations above 0.15 mg/L, likely a result of over-fitting to the training data during the stacking process.

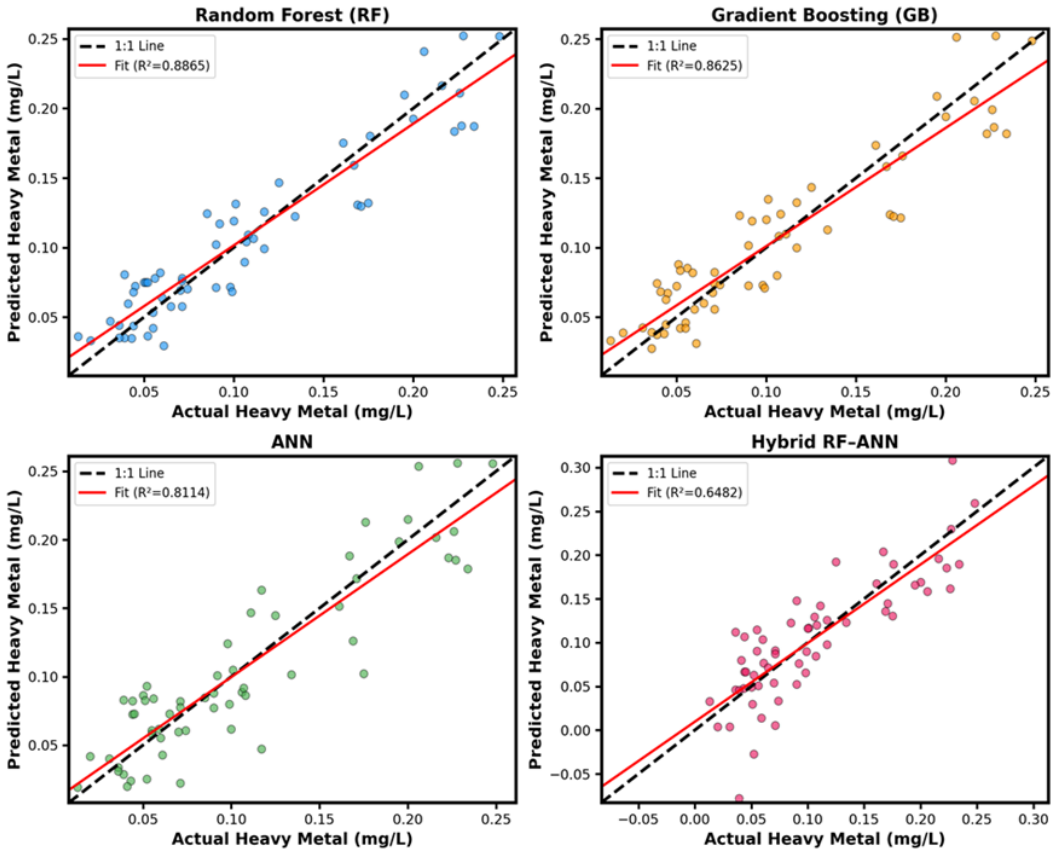

Figure 3: *Actual vs. Predicted Heavy Metal Concentration for All machine Learning*

Figure 3 shows a 2 x 2 grid of scatter plots, with the x-axis showing the true concentration of the heavy metals and the y-axis the model predictions. If the model is perfect, then all of the points will lie on the dashed black 1:1 line: the model prediction is equal to the true concentration, whatever that concentration is. The red regression line through the scatter of each plot gives an indication of bias: if the red line is close to and overlays the dashed line, then the model is fit and unbiased; if it is away from the dashed line (sloping upward or downward) then the model is biased. Let's first look at the RF panel. The fit is good, the red regression line is almost identical to the dashed 1:1 line, and we do not see any increase in scatter at high concentrations, which means that the model is equally good at predicting the concentration of relatively unpolluted water (less than 0.05 mg/L) as it is good at predicting the concentration of moderately polluted water (above 0.20 mg/L). This feature is very desirable for environmental applications because many real-world water quality data sets are poorly balanced towards low concentrations and a model that fits that part of the concentration range is not very useful when a pollution event occurs and concentrations increase. The GB panel shows a similar pattern as RF, but there is more scatter at high concentrations. A closer look reveals that several points at the highest concentrations are slightly below the 1:1 line: the model slightly underestimates the concentration. This is typical of gradient boosting with heavy regularization (low learning rate, shallow trees), the model scales down the correction at each step to avoid overfitting and this can cause a slight bias towards underestimating the outliers.

This is the safe approach for a monitoring application, the model would not miss a pollution event but may underestimate it (which is likely to be better than overestimating).The ANN panel shows a much larger point cloud with a larger spread around the 1:1 line in the middle of the concentration range (0.08-0.18 mg/L). That's a sign that the neural network has not learned the smooth monotonic relationship between pH and heavy metal concentration very well from the training set of 240 samples, it would approach the tree-based panels with a larger amount of training data. Finally, the Hybrid RF-ANN panel is most visually appealing: the regression line is bowed (down) and underestimates concentrations above approximately 0.15 mg/L. This is a common behavior of a stacked model in which the secondary learner has overfit the fit residuals (residuals from the primary learner's fit) of the primary learner, it has learned to fit the noise (and some of the signal) of the primary learner's fit (and

this issue could be substantially remedied by a larger experimental dataset, which would provide more diverse training samples to the stacking learner).

3.4 Feature Importance and Mechanical Interpretation

The feature importance analysis (Figure 4) reveals a mechanistically plausible interpretability of the predictor features. pH has the highest rank (MDI = 0.907), which is in line with the results of its overwhelming correlation with the targeted heavy metal that were reported in Section 3.1. The dominant role of pH is rooted in fundamental aqueous geochemistry. As pH decreases, the rising concentration of H^+ ions causes proton competition with cations for binding sites on hydroxide, carbonate, and organic ligands. This drives the dissolution of metal hydroxide and carbonate minerals phases and promotes desorption of metal ions from surfaces, releasing previously immobilized metals into dissolve form. Consequently, even a modest drop in pH across the experiment range (5.52-7.05) produces a measurable increase I free metal ion activity. From a model behavior perspective, the RF algorithm has effectively learned this thermodynamic speciation relationship, pH-based node splits alone account for 90.7 % of total impurity reduction, meaning pH carries sufficient information to partition sample into low-and high-contamination categories, with the remaining features contributing secondary refinements within those partitions. This is important for designing monitoring systems: the pH sensor is not a "me too" quality control feature but the most informative sensor signal that can be derived from the sensor suite and, as such, its temporal resolution and frequency of calibration should be prioritized accordingly.

However, electrical conductivity ranked second (MDI = 0.034), followed by temperature (MDI = 0.018), apparent resistance (MDI = 0.016), current (MDI = 0.011), voltage (MDI = 0.008), and power (MDI = 0.007). The ranking of the permutation importance confirms this ordering for the first two features (pH and conductivity), and shows that the MDI ranking is not merely a consequence of the large number of features used. The relatively high importance of the engineered resistance feature (rank 4 in MDI) justifies the feature engineering step, the individual features (voltage and current) by themselves each contain very little information about the response, but their ratio (apparent resistance) contains information on the ionic strength and electrode kinetics of the sample matrix that is not encoded in the individual features. This finding validates the theoretical motivation for introducing derived electrochemical parameters and implies that future sensor designs could profit from the read-out of impedance-based features, in addition to classical voltammetric measurements [33].

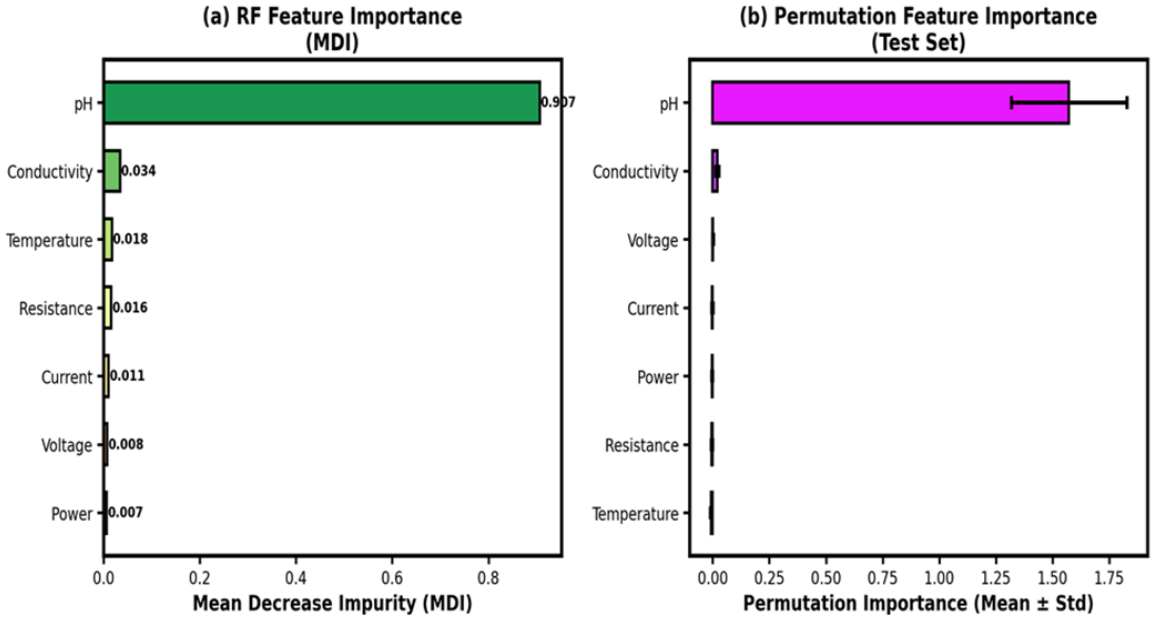

Figure 4: *Feature Importance Analysis: Identifying Key Electrochemical and Physiochemical Predictors*

In Figure 4, we display the feature importance analysis in two panels. The chart on the left (a) shows the MDI as horizontal bars with a yellow to green color gradient. The most obvious thing to note is that pH towers over all of the other variables, its bar reaches 0.907 on the importance scale, while every other variable is crammed into the sliver below 0.035. The pH bar is so much longer than all other bars that it seems almost like a whole other bar chart, but this is the right and physically reasonable result. The important thing to understand about the MDI score is that, among all of the 200 decision trees and the millions of node splits they all make, the split based on pH value is responsible for 90.7% of the error reduction. There is simply nothing like it. Scanning the rest of the MDI ranking, the variables conductivity (0.034) and temperature (0.018) come second and third, with the engineered resistance feature (0.016) in fourth, note that the latter is computed from the raw current (0.011), voltage (0.008), and power (0.007) variables, which come fourth, fifth, and sixth in the ranking. The significance of this sequence is that V/I contains information about the sample's electrochemical status that is not contained either in V or I. A sensor unit that simply reports the raw voltammetric signals would provide less predictive power than one that also calculates (and reports) the apparent impedance of the electrochemical cell. The right-hand panel (b), showing the permutation importance, with ±1 standard deviation plotted for 30 repetitions of the shuffling, confirms the general ranking while bringing a dose of realism to the ranking. Permutation importance operates by deliberately scrambling the information in one variable at a time (by random shuffling of the variable's values for the test set samples) and then assessing the degradation in model performance. If a variable carries valuable (non-redundant) information, shuffling it will have a large effect, if the information is redundant with other variables, it will not. The order of pH and conductivity in panel (b) is consistent with the MDI rankings in panel (a), suggesting that the latter are not unduly influenced by high cardinality of features the two approaches agree. The small error bars on pH's importance confirm that this ranking is not a fluke outcome of a single shuffle but is consistent over multiple shuffles. Both panels are thus strongly consistent in their physical reality, pH sensor performance is the most important factor affecting the accuracy of the framework, and the most money should be spent on pH sensor calibration.

3.5 Comparison with Published Literature

The performance of the RF model (R^2 = 0.887) is consistent with a variety of published machine learning models used in predicting heavy metal concentrations in water systems. Ahmed et al. [34] achieved R^2 = 0.81-0.86 with RF models for arsenic prediction in groundwater from Bangladesh, using geochemical data as predictors but no electrochemical data. Bui et al. [35] reported R^2 = 0.79 for an ANN model prediction of heavy metals in Vietnamese rivers, but with model performance noted to depend on the training dataset. Zhu et al. [36] achieved R^2 = 0.84 with gradient boosted trees for multi-metal prediction from hyperspectral remote sensing, but remote sensing methods are limited to surface water light scattering and cannot detect dissolved metals at environmentally relevant concentrations without proxy relationships with particulate matter. The key innovation of the present approach relative to these examples is the inclusion of signals from the electrochemical sensors, providing first-principles physicochemical information about the activity of the ions that cannot be obtained from optical or geochemical covariate sets. This is a significant innovation that is encapsulated in the high R^2 obtained with a relatively small (240 observations) training set.

3.6 Environmental Relevance and Monitoring

The significance of R^2 = 0.887 and RMSE = 0.022 mg/L needs to be considered relative to drinking water monitoring thresholds. The WHO maximum permissible level (MCL) for lead in drinking water is 0.010 mg/L, at the average concentration in the test set (0.107 mg/L), the RMSE of the RF model is around 21% of the mean and less than two MCL units, implying that the model will correctly classify the vast majority of samples as above or below the MCLs. This classification capability forms the basis for the potential application of the framework in a two-tiered monitoring strategy that involves the use of a rapid, inexpensive ML-based pre-screen to classify high-priority samples for follow-up laboratory analysis, thus optimizing the allocation of analytical effort where it can provide the most risk reduction [37]. Coupling to IoT sensor networks where the electrochemically and physiochemically relevant parameters measured in this study can be communicated in near real-time to cloud-based inference platforms will also contribute to the practical value of the framework [38]. It should be noted that these represent prospective development scenarios, validation under real field conditions remains a subject for future investigation.

4.0 Conclusions

This work has shown that a transparent machine learning model, incorporating the response from electrochemical sensors with traditional physico-chemical water quality parameters and engineered features, can accurately and interpretably predict dissolved heavy metal concentrations from an experimental data set. The major findings are the following: (i) The ensemble RF model provided the best generalization on the test set (R^2 = 0.887, RMSE = 0.022 mg/L, MAE = 0.018 mg/L) in comparison to GB, single ANN, and Hybrid RF-ANN models. (ii) Feature importance analysis (considering both MDI and permutation importance) confirmed pH as the overwhelmingly most important predictor of dissolved heavy metal concentration (MDI weight 0.907), followed by electrical conductivity. This is in perfect agreement with the chemistry of solution speciation of metal ions as a function of pH. (iii) The engineered electrical resistance feature (R = V/I) was a significant predictor, beyond the raw electrochemical features, demonstrating the value of physico-electrochemically inspired feature engineering to enhance the information content of chemical sensor data. (iv) The stacking (RF-ANN) model did not perform better than the single RF model on this dataset, probably because of the limited number of samples (n = 240) available for training. This strategy should be further tested on larger training sets. More broadly, the dataset size of 300 samples represents a recognized limitation of the present study, all reported metrics should be interpreted with this constraint in mind, and the generalization of the framework to diverse water matrices requires validation on substantially larger datasets in future work. (v) The model is explainable, efficient, and computationally lightweight, with potential suitability for future deployment on IoT-enabled sensor platforms, however validation under a real field condition is required before operational use in automated heavy metal monitoring of freshwater system can be recommended.

5.0 Future Work

There are number of avenues for future research. First, the experimental data set should be increased to several thousand observations, including measurements from a range of different water matrices (groundwater, treated effluent, and drinking water distribution systems) for more robust models and to test transfer learning across different monitoring sites. Second, combining miniaturized impedance spectroscopy with voltammetric determinations would enhance electrochemical features for more accurate predictions in complex matrices. Third, the implementation of the framework on embedded microcontroller systems (e.g., Raspberry Pi, Arduino Portenta) and LoRaWAN/NB-IoT communication infrastructure would confirm the real-world performance of the sensor system in terms of latency and data quality. Fourth, multi-target prediction (prediction of concentrations of multiple metal species) would considerably enhance the value of the sensor system. Finally, it would be interesting to explore the integration of the framework into digital twin technology for urban water systems for risk assessment at the catchment level [39].

Funding: No specific funding was received from any funding agency in the public, commercial or not-for-profit sectors to carry out this research.

Conflict of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Data Availability: The experimental data in this paper are available from the corresponding author on request.

Author Contributions: The corresponding author conceived the idea, collected and analysed the data, and wrote and edited the manuscript.

Reference

[1] Vardhan, K.H., Kumar, P.S., & Panda, R.C. (2019). Review on heavy metal pollution, toxicity, and remedial action: Current status and future directions. Journal of Molecular Liquids, 290, 111197. https://doi.org/10.1016/j.molliq.2019.111197
[2] Fu, Z., & Xi, S. (2020). How heavy metals affect human metabolism. Toxicology Mechanisms and Methods, 30(3), 167–176. https://doi.org/10.1080/15376516.2019.1701594
[3] Balali-Mood, M., Naseri, K., Tahergorabi, Z., Khazdair, M.R., & Sadeghi, M. (2021). Mechanism of action of five toxic metals: mercury, lead, chromium, cadmium, and arsenic. Frontiers in Pharmacology, 12, 643971. https://doi.org/10.3389/fphar.2021.643971

[4] WHO (2022). Guidelines for Drinking-water Quality (4th ed., including 1st and 2nd addenda). World Health Organization, Geneva. ISBN 978-92-4-004506-4.
[5] Tchounwou, P.B., Yedjou, C.G., Patlolla, A.K., & Sutton, D.J. (2012). Environmental heavy metal toxicity. Experientia Supplementum, 101, 133–164. https://doi.org/10.1007/978-3-7643-8340-4_6
[6] Prüss-Ustün, A., Wolf, J., Corvalán, C., Bos, R., & Neira, M. (2016). Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks. World Health Organization, Geneva.
[7] Fashola, M.O., Ngole-Jeme, V.M., & Babalola, O.O. (2016). Gold mine pollution and effects on microorganisms. International Journal of Environmental Research and Public Health, 13(11), 1047.
[8] European Parliament and Council (2000). Directive 2000/60/EC on the establishment of a framework for Community action in the field of water policy. Official Journal of the European Communities, L 327, 1-73.
[9] Mandal, P., Upadhyay, R., & Hasan, A. (2010). Spatial and temporal variation of water quality of the Yamuna River in Delhi, India. Environmental Monitoring and Assessment, 170(1), 661–670. https://doi.org/10.1007/s10661-009-1261-8
[10] Noh, H., Chung, E., Kim, Y., Kim, S.D., & Park, J. (2016). Water quality monitoring network in the Han River basin. Water, 8(12), 551. https://doi.org/10.3390/w8120551
[11] Rode, M., et al. (2016). High-frequency monitoring in the stream: The wave of the future. Environmental Science & Technology, 50(19), 10297–10307. https://doi.org/10.1021/acs.est.6b02155
[12] Monzó, P., et al. (2020). Determination of trace metals by voltammetric methods. TrAC Trends in Analytical Chemistry, 130, 115955. https://doi.org/10.1016/j.trac.2020.115955
[13] Aguilar-Lira, G.Y., et al. (2019). Recent advances in stripping techniques for determination of heavy metals in environmental water samples by using portable systems and screen-printed electrodes. Analytical Methods, 11(14), 1874–1885. https://doi.org/10.1039/C9AY00224C
[14] Lu, Y., Liang, X., Niyungeko, C., Zhou, J., Xu, J., & Tian, G. (2018). The detection and identification of environmental heavy metal ions by voltammetry. Talanta, 178, 324–338. https://doi.org/10.1016/j.talanta.2017.08.033
[15] Claverie, R., Nguyen, A., Phalyvong, K., Sébès, L., Wiest, L., & Vulliet, E. (2020). Effect of fouling on electrochemical sensing. Electrochimica Acta, 362, 137063. https://doi.org/10.1016/j.electacta.2020.137063
[16] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., & Ye, L. (2022). Machine learning for water quality assessment: a review. Eco-Environment & Health, 1(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001
[17] Haghiabi, A.H., Nasrolahi, A.H., & Parsaie, A. (2018). Predicting water quality using machine learning. Water Quality Research Journal, 53(1), 3–13. https://doi.org/10.2166/wqrj.2018.025
[18] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Enhancing prediction of water quality indices by new hybrid machine learning models. Science of the Total Environment, 721, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612
[19] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[20] Grbović, F., Bajić, D., Petrović, N., & Zorić, A. (2023). Machine learning and integrated electrochemical sensors for real-time water quality monitoring. Sensors, 23(4), 2219. https://doi.org/10.3390/s23042219
[21] Khullar, S., & Singh, N. (2021). River water quality modeling using a deep learning Bi-LSTM approach. Environmental Science and Pollution Research, 29(9), 12885–12898. https://doi.org/10.1007/s11356-021-14875-6
[22] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). Self-published. https://christophm.github.io/interpretable-ml-book/
[23] Economou, A., & Fielden, P.R. (2003). Mercury film electrodes: developments, trends, and potentialities for electroanalysis. Analyst, 128(3), 205–212. https://doi.org/10.1039/b212095h
[24] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer, New York. https://doi.org/10.1007/978-0-387-84858-7
[25] Barsoukov, E., & Macdonald, J.R. (2018). Impedance Spectroscopy: Theory, Experiment, and Applications (3rd ed). Wiley, New Jersey. https://doi.org/10.1002/9781119381860
[26] Liaw, A., & Wiener, M. (2002). RandomForest: a method of classification and regression. R News, 2(3), 18–22.
[27] Chen, T., & Guestrin, C. (2016). XGBoost: Scalable, portable, and distributed gradient boosting. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.
[28] Wolpert, D.H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[29] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134
[30] Stumm, W., & Morgan, J.J. (1996). Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters (3rd ed.). Wiley, New York.
[31] Olías, M., Cánovas, C.R., Nieto, J.M., & Sarmiento, A.M. (2004). Assessment of dissolved load carried by the Tinto and Odiel rivers (Southwest Spain). Applied Geochemistry, 19(10), 1733–1745. https://doi.org/10.1016/j.apgeochem.2004.04.001
[32] Chai, T., & Draxler, R.R. (2014). Why avoid root mean square error (RMSE) in the literature: a case for RMSE or mean absolute error (MAE). Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014
[33] Randviir, E.P., & Banks, C.E. (2015). Electrochemical impedance spectroscopy: An overview of bioanalytical applications. Analytical Methods, 5(5), 1098–1115. https://doi.org/10.1039/C2AY26119E
[34] Ahmed, K.R., Paul, S.C., Jannat, M., & Islam, M.M. (2021). Machine learning-based prediction of arsenic contamination in groundwater. Chemosphere, 282, 131019. https://doi.org/10.1016/j.chemosphere.2021.131019
[35] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Optimal prediction of water quality indices by new hybrid machine learning approaches. Science of the Total Environment, 721, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612
[36] Zhu, W., et al. (2020). Machine learning for estimating chromophoric dissolved organic matter in inland lakes. Remote Sensing of Environment, 241, 111740. https://doi.org/10.1016/j.rse.2020.111740
[37] Behmel, S., Damour, M., Ludwig, R., & Rodriguez, M.J. (2016). A review and outlook on water quality monitoring strategies. Science of the Total Environment, 571, 1312–1329. https://doi.org/10.1016/j.scitotenv.2016.06.235
[38] Adu-Manu, K.S., et al. (2017). A review of water quality monitoring with wireless sensor networks. ACM Transactions on Sensor Networks, 13(3), 1–40. https://doi.org/10.1145/3005719
[39] Mooij, P.R., & Robrecht, G.H. (2020). Digital twins for smart urban water management. Urban Water Journal, 17(3), 285–296. https://doi.org/10.1080/1573062X.2020.1726412