sustainability

MDPI

Article

Machine Learning-Based Prediction and Interpretation of Aluminum and Iron Coagulant Dosages in Drinking Water Treatment

Qian-Kun Lian [1,2], Zhu-Jun Feng [1,2], Zhi-Yuan Xu [1,2], Chen-Ming Xue [1,2], Yan-Ting Huang [1,2], Xiao-mao Wang [3], Yuanquan Xu [1,2,*] and Wen-Jie Fu [1,2,*]

[1] Guangxi Key Laboratory of Environmental Processes and Remediation in Ecologically Fragile Regions, Guangxi Normal University, Guilin 541004, China; lianqiankun2025@163.com (Q.-K.L.); xzy7039@163.com (Z.-Y.X.)
[2] College of Environment and Resources, Guangxi Normal University, Guilin 541004, China
[3] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China; wangxiaomao@tsinghua.edu.cn
* Correspondence: yqxu@gxnu.edu.cn (Y.X.); fuwj@gxnu.edu.cn (W.-J.F.)

Abstract

Accurate and responsive coagulant dosage control is essential for maintaining drinking water treatment performance. In this study, the multi-WTP ICR database was used to develop machine learning (ML) models for Al- and Fe-based coagulant dosage prediction. Four ML algorithms, including CatBoost, neural network (NN), random forest (RF), and support vector regression (SVR), were systematically evaluated, and SHapley Additive exPlanations (SHAP) were employed to interpret model predictions. CatBoost and NN achieved comparable predictive performance for both Al and Fe coagulant dosages, followed by RF, whereas SVR showed the lowest predictive accuracy. SHAP analysis identified TOC, UV_{254}, alkalinity, and total hardness as the major contributing features in both models. Preliminary testing using jar tests and data from 12 WTPs reported in published studies yielded MAPE values of 5.16% and 21.09%, respectively, for the CatBoost model predicting Al-based coagulant. An online platform integrating dosage prediction and SHAP interpretation was further developed to provide an interpretable, data-driven approach for coagulant dosage estimation and support more responsive dosing decisions in drinking water treatment.

Keywords: drinking water treatment; coagulant dosage; machine learning; CatBoost; SHAP

Academic Editor: Jun Hou

Received: 28 September 2026
Revised: 25 October 2026
Accepted: 2 November 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Coagulation plays a central role in water treatment processes as it facilitates the removal of suspended particles, colloids, and a fraction of dissolved organic matter, thereby affecting the loading of subsequent sedimentation and filtration processes and the quality of treated water [1]. The selection of an appropriate coagulant dosage is crucial for effective operation of coagulation, which depends on both influent water quality and the desired effluent quality. Insufficient dosage results in incomplete destabilization of colloidal particles and elevated effluent turbidity [2]. In contrast, excessive dosing leads to unnecessary chemical consumption, increased sludge production, and additional water

quality concerns, such as elevated residual aluminum (Al) or iron (Fe) concentrations [3]. Therefore, accurate and responsive coagulant dosage control is important not only for maintaining treatment performance but also for improving resource efficiency and reducing the environmental footprint of water treatment operations.

Currently, most water treatment plants (WTPs) still rely on operator experience and jar tests to determine coagulant dosages. The former is highly susceptible to operator-to-operator variability, whereas the latter increases labor and experimental costs and cannot facilitate real-time adjustment in response to fluctuations in influent water quality [4]. With the increasing deployment of online sensors and data acquisition systems, some modern WTPs have begun to adopt automated dosing systems based on flow rate and several water quality parameters, such as influent temperature and pH. However, these methods often exhibit limited predictive performance because they are primarily based on linear models [5]. Machine learning (ML) algorithms offer an alternative for addressing this limitation, as they can capture complex nonlinear relationships and interactions among multiple variables involved in water treatment processes [6,7].

Preliminary efforts have been made to apply ML to coagulant dosage prediction. Early studies mainly employed support vector regression (SVR) and artificial neural networks (ANNs) for aluminum-based coagulant dosage prediction, with datasets typically comprising several hundred records [8–11]. Zhang et al. reported limited transferability of SVM models across WTPs. Using influent pH, temperature, and turbidity as input features, their SVM model trained on data from a large-scale WTP achieved an R^2 of only approximately 0.4 when applied to a small-scale WTP [8]. Wu et al. investigated the effects of input features on ANN prediction accuracy and found that the coagulant dosage on the preceding day contributed substantially to model performance [10]. However, heavy reliance on the previous day's dosage, particularly when predicted dosages are recursively used as inputs, may introduce error propagation over successive prediction steps.

As monitoring data accumulate, recent studies have begun to employ ML algorithms with stronger nonlinear modeling capabilities to capture the complex relationships between water quality parameters and coagulant dosage [12,13]. Wang et al. evaluated the performance of recurrent neural networks (RNNs) and random forests (RFs) for aluminum-based coagulant dosage prediction using a dataset containing more than 2000 observations. Their results demonstrated that the performance of both models was strongly influenced by feature selection. Using temperature, flow rate, and influent turbidity as the baseline features and incorporating different combinations of dissolved oxygen, pH, and oxygen consumption, the RF models yielded RMSE values of 1.09–1.87 mg/L and R^2 values of 0.17–0.58, whereas the RNN models generally performed better, with RMSE and R^2 values ranging from 0.60 to 1.47 mg/L and from 0.24 to 0.87, respectively [14]. In addition, because influent water quality exhibits distinct clustering patterns across seasons and hydrological conditions, hybrid cluster-Adaptive Neuro Fuzzy Inference System (ANFIS) approaches have been proposed, in which clustering is performed before constructing ANFIS models [9,11]. Using 8755 hourly records from a Korean WTP, Kim et al. trained a k-means-ANFIS model with influent and effluent turbidity, water temperature, pH, conductivity, and flow as inputs, achieving an RMSE of 1.848 mg/L and an R^2 of 0.9259, compared with 2.075 mg/L and 0.9066 for standalone ANFIS and 2.2272 mg/L and 0.8924 for ANN [9]. Collectively, these studies demonstrate the potential of ML for automated water treatment. However, most of them rely on data from individual WTPs, constraining their generalizability and transferability. Insufficient model interpretability makes it difficult to identify the key factors influencing coagulant dosage. Moreover, few studies have progressed beyond model development and validation toward practical, open-source application for coagulant dosing.

To address these limitations, this study used the multi-WTP ICR database to develop Al- and Fe-based coagulant dosage prediction models using four ML algorithms, including CatBoost, RF, NN, and SVR. Their performance was systematically evaluated, with SHapley Additive exPlanations (SHAP) employed to identify the key factors influencing coagulant dosage. The Al dosage model was further tested through preliminary jar tests and data from 12 WTPs reported in published studies. Finally, the best-performing models and SHAP analysis were integrated into an open-source platform (icr-coagulant.streamlit.app), providing an interpretable, data-driven tool to support coagulant dosing decisions in WTPs.

2. Materials and Methods

2.1. Data Collection and Processing

The data used in this study were obtained from the U.S. EPA Information Collection Rule (ICR) drinking water treatment database, which contains water quality and operational records from 296 WTPs across the United States collected between July 1997 and December 1998 [15,16]. Influent and effluent water quality parameters, WTP types, and coagulant dosages were extracted from the database. It should be noted that the coagulant dosage values recorded in the ICR database represent historical dosages applied during routine WTP operation and are not necessarily optimal dosages. Based on coagulant types, the data were divided into two independent datasets, Al-based and Fe-based coagulant datasets (Al-CD and Fe-CD). The former comprised 1591 records, while the latter comprised 487 records.

Data processing involved five steps, including removal of records with missing values, category merging, one-hot encoding, stratified random splitting of the dataset into training and test sets at a ratio of 8:2 within each WTP type, and Z-score standardization. Category merging was adopted because some treatment categories had very few samples. Categories with small sample sizes were therefore merged based on the similarity of their input features (see Text S1 for details). A sensitivity analysis was also conducted to evaluate the effect of this merging strategy on model performance. The results showed that the merging strategy had a negligible effect on predictive performance (Table S11). CatBoost was trained using data after missing-value handling and category merging, without one-hot encoding or Z-score standardization. Data for the RF model did not require Z-score standardization, whereas the SVR and NN models were trained using data after Z-score standardization. Detailed procedures are provided in Text S1.

2.2. Model Development

The modeling task was formulated as a supervised regression problem, with the historical coagulant dosage recorded in the ICR database as the prediction target and water quality and operational parameters as input features. Four ML algorithms, including CatBoost [17], NN [18], RF [19], and SVR [20], were selected here. In addition, multiple linear regression (MLR) was included as a baseline for comparison. For both Al-CD and Fe-CD, the models used nine continuous variables: influent pH, alkalinity (mg/L as $CaCO_3$), turbidity (NTU), temperature (°C), total hardness (mg/L as $CaCO_3$), total organic carbon (TOC, mg/L as C), ultraviolet absorbance at 254 nm (UV_{254}, cm^{-1}), influent flow rate (MGD), and effluent turbidity (NTU). Al-CD included plant type and disinfection type, yielding 11 input features. Fe-CD further includes water source type, resulting in 12 input features. Although the actual effluent turbidity is unavailable before dosing, the target effluent turbidity can be specified in advance based on drinking water quality requirements and treatment objectives. It was therefore included as an input

feature to enable the model to predict coagulant dosage according to the desired effluent quality.

The NN model was implemented as a fully connected neural network using mean squared error (MSE) as the loss function. Hyperparameters were optimized using grid search, including the number of hidden layers, number of neurons per layer, dropout rate, weight decay, optimizer type, batch size, and learning rate. CatBoost was trained using root mean squared error (RMSE) as the loss function, with the tree depth, L2 regularization coefficient for leaf values, random strength for split scoring, minimum number of samples per leaf, subsampling ratio, and learning rate optimized. For RF, the main hyperparameters included the number of trees, maximum tree depth, minimum number of samples per leaf, minimum number of samples required for node splitting, and number of candidate features considered at each split. SVR employed the ε-insensitive loss function, with the kernel type, penalty parameter, and kernel coefficient optimized.

All models were trained and compared using the same training–test split. Model performance was evaluated using the coefficient of determination (R^2), RMSE, and mean absolute error (MAE). Model development was conducted in Python 3.11 using CatBoost 1.2.10, PyTorch 2.14.0, and scikit-learn 1.8.0. Detailed training procedures are provided in Text S2. The optimized hyperparameters for each model are listed in Tables S4–S7, and the corresponding learning curves are presented in Figure S8.

2.3. Shap Analysis

SHAP was used to quantify the contribution of each input feature to model predictions [21]. Based on Shapley values from cooperative game theory, SHAP treats input features as players and allocates the difference between a prediction and its baseline value among them. For a sample x, the SHAP value of feature i is defined as the weighted average of the changes in model output resulting from adding that feature to different subsets of features:

$$\varphi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[v_x(S \cup \{i\}) - v_x(S)\right]$$

Here, F is the set of all input features, M is the total number of features, S is any subset that excludes feature i, and $|S|$ is the number of features in that subset. The quantity $v_x(S)$ denotes the expected model output when the information for features in S is retained for sample x and the remaining features are handled according to the explainer settings. The weighting coefficient accounts for the different possible orders in which features can be added. SHAP satisfies the local accuracy (additivity) property:

$$f(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i(x), \varphi_0 = v_x(\emptyset)$$

Here, $f(x)$ is the model prediction and φ_0 is the baseline output. A positive SHAP value indicates that a feature increases the predicted dosage relative to the baseline, whereas a negative value indicates a decrease. The absolute SHAP value represents the magnitude of the feature's contribution. TreeExplainer was applied to the trained Al-CD and Fe-CD CatBoost models to calculate SHAP values for each feature on the respective test sets [22].

2.4. Coagulant Dosing Experiments with Actual Plant Influent

Jar coagulation tests were conducted using raw water collected from a WTP to provide a preliminary assessment of model performance under plant-relevant conditions. The treatment process consists of coagulation, sedimentation, and chlorine disinfection, which corresponds to the conventional type in the ICR database. Six influent water samples were randomly collected on different dates in June 2026. Three samples were tested with an alum dosage of 2.5 mg/L as Al, while the other three were tested at 3.0 mg/L as Al. For the jar tests, rapid mixing was first conducted at 200 rpm for 4 min, followed by slow mixing at 40 rpm for 15 min to promote flocculation, and then sedimentation for 60 min. Influent turbidity, pH, water temperature, alkalinity, TOC, UV_{254}, and total hardness were measured, together with the turbidity of the supernatant after sedimentation. Detailed analytical methods for the water quality parameters are provided in Text S3. The measured parameters and other required inputs, including the plant's monthly average influent flow rate (63.4 MGD), WTP type (CONV), disinfection type (Cl2), and supernatant turbidity after 60 min of settling (used as the effluent turbidity input), were entered into the trained CatBoost Al dosage model to predict the coagulant dosages. Model prediction accuracy was evaluated by comparing the absolute and relative errors between the applied and predicted dosages.

3. Results and Discussion

3.1. Data Characteristics and Correlation Analysis

Figure 1 and Table S1 present the distributions of water quality parameters in Al-CD and Fe-CD, with the ranges discussed in the main text representing the interquartile ranges (25th–75th percentiles) and those in Table S1 indicating the minimum and maximum values. The Al-CD covered influent waters with pH values of 7.30–8.10 and relatively broad ranges of alkalinity (28.00–109.00 mg/L as $CaCO_3$) and total hardness (40.00–141.00 mg/L as $CaCO_3$). Influent TOC and UV_{254} ranged from 1.95 to 3.60 mg/L C and 0.042 to 0.105 cm^{-1}, respectively, whereas turbidity and temperature varied from 1.50 to 8.60 NTU and 11.00 to 22.00 °C, respectively. Under these water quality conditions, approximately 60.2% of the samples had coagulant dosages in the range of 0–2 mg/L Al, with a mean and median dosage of 2.33 and 1.60 mg/L Al, respectively.

Compared with Al-CD, Fe-CD was characterized by slightly higher alkalinity and hardness, with values of 69.50–139.00 and 93.00–208.00 mg/L as $CaCO_3$, respectively, which may be among the factors associated with the use of Fe-based coagulants, as higher alkalinity provides greater buffering capacity against the pH decrease associated with iron salt hydrolysis [15]. The influent pH (7.50–8.07), TOC (2.50–4.80 mg/L C), UV_{254} (0.056–0.111 cm^{-1}), turbidity (1.88–11.10 NTU), and temperature (13.65–24.65 °C) were also within relatively narrow ranges. The mean and median Fe dosages were 5.58 and 3.30 mg/L Fe, respectively.

Both Al and Fe dosages exhibited right-skewed distributions, with the mean values exceeding the corresponding medians. This pattern may reflect relatively stable coagulant demand under normal operating conditions, whereas substantially higher dosages may be required when raw-water quality deteriorates or unusual water quality events occur to maintain treated-water quality, resulting in a number of high-dose observations.

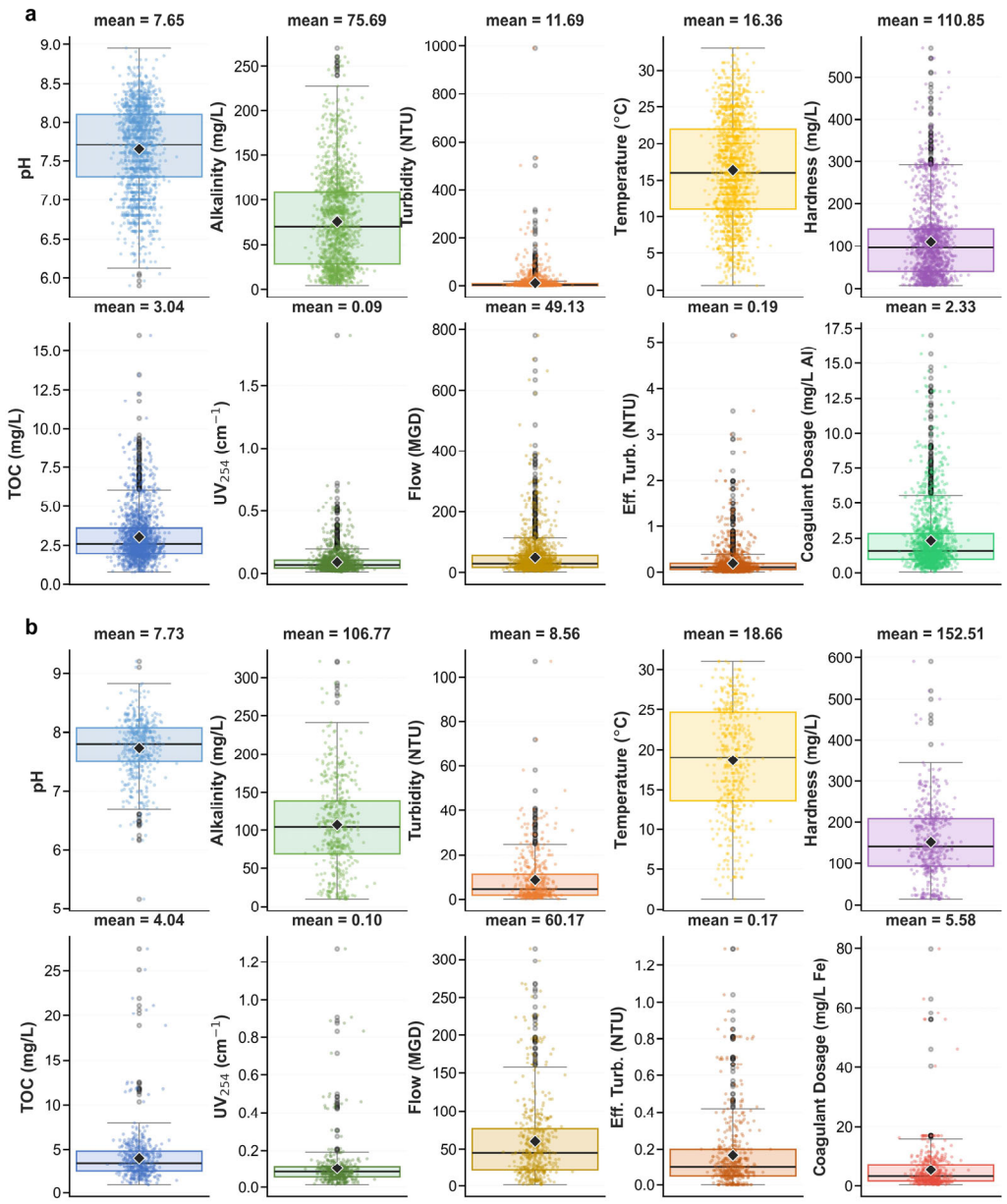


Figure 1. Distribution of water quality parameters in (**a**) Al-CD and (**b**) Fe-CD. The boxes represent the interquartile range (IQR), and the whiskers extend to the data points within 1.5× IQR. Colored dots represent the randomly jittered individual observations, while black diamonds indicate the mean values. Mean values are annotated above the corresponding boxes.

Correlations among water quality parameters and coagulant dosage were further examined (Figure 2a,b). In Al-CD, pH, alkalinity, and hardness were all significantly positively correlated with one another, potentially reflecting the combined influence of carbonate chemistry and source-water geochemical processes. The weathering of carbonate minerals can simultaneously release Ca^{2+}/Mg^{2+} and HCO_3^-, thereby contributing to the strong association between water hardness and alkalinity. Meanwhile, the carbonate buffering system plays an important role in regulating acid–base equilibria, which may contribute to the observed correlations of pH with alkalinity and hardness. As important indicators of the abundance and compositional characteristics of dissolved organic matter, TOC and UV_{254} were strongly and positively correlated ($r = 0.79$, $p < 0.001$),

indicating that higher organic carbon concentrations were generally accompanied by greater concentrations of UV-absorbing organic compounds. Among the water quality parameters, TOC, UV_{254}, and influent turbidity were all significantly and positively correlated with coagulant dosage, with correlation coefficients of 0.509, 0.538, and 0.402, respectively. Higher turbidity generally indicates increased suspended and colloidal particle loads in raw water, which may increase coagulant demand to promote particle aggregation through charge neutralization, adsorption, and sweep flocculation. Similarly, elevated TOC and UV_{254} indicate a higher organic matter load, which also requires higher coagulant dosages to achieve effective removal.

A similar correlation pattern was observed in Fe-CD. TOC and UV_{254} showed a strong positive correlation (r = 0.87, $p < 0.001$), while TOC, UV_{254}, and influent turbidity were significantly and positively correlated with Fe-based coagulant dosage, with coefficients of 0.301, 0.325, and 0.389, respectively. These results suggest that organic matter and particulate loads in raw water represent common factors associated with the demand for both types of coagulants. Notably, the correlations of flow rate with coagulant dosage and organic matter indicators showed contrasting patterns between the two datasets. In Al-CD, flow rate was negatively correlated with coagulant dosage, TOC, and UV_{254}, whereas positive correlations were observed in Fe-CD. The correlation between flow rate and coagulant dosage may therefore not reflect a direct effect of flow rate on coagulant demand, but rather the coupled variation between flow rate and raw-water quality under different source-water and plant operating conditions.

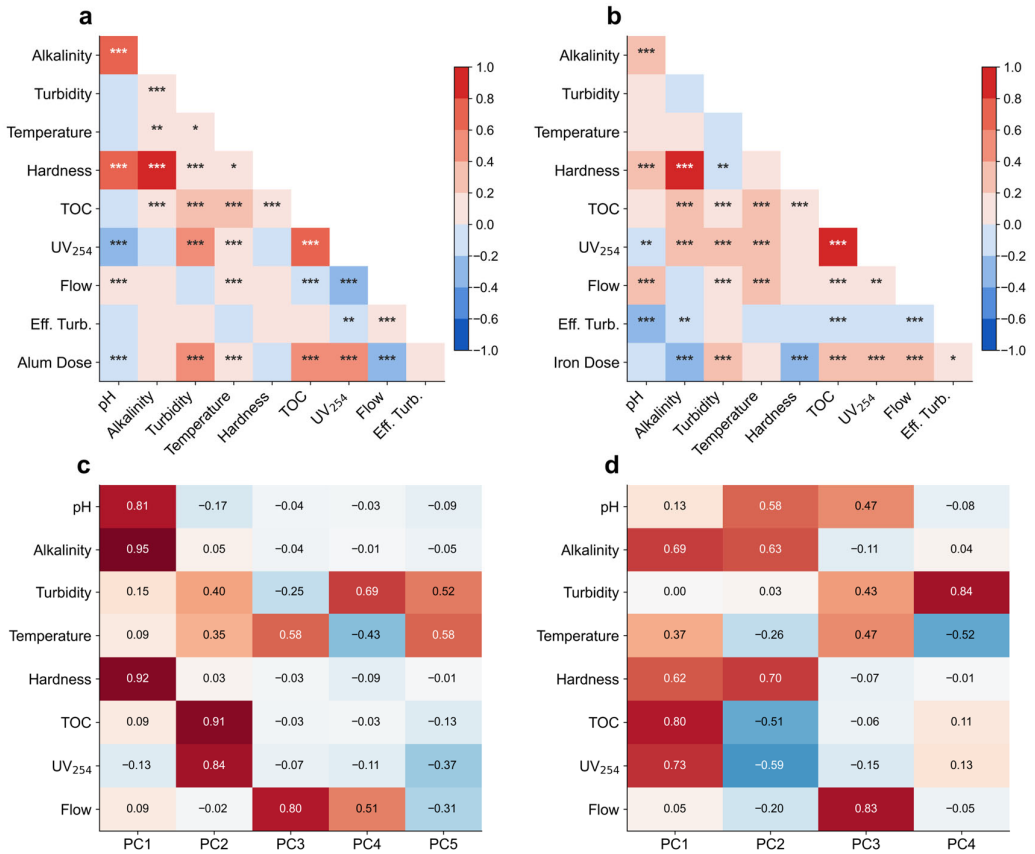


Figure 2. Correlation patterns and PCA loadings of water quality parameters in Al-CD and Fe-CD. (**a**) Spearman correlation heatmap and (**c**) PCA loadings for Al-CD; (**b**) Spearman correlation heatmap and (**d**) PCA loadings for Fe-CD. Significance levels are indicated by * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

To further characterize the overall variation patterns of water quality parameters in Al-CD and Fe-CD, principal component analysis (PCA) was performed on standardized continuous influent variables (Figures 2c,d and S5). In Al-CD, the first five principal components (PCs) accounted for a cumulative 89.5% of the total variance. PC1 was primarily associated with pH, alkalinity, and hardness, with loadings of 0.81, 0.95, and 0.92, respectively, suggesting that this component mainly represents the mineralization and carbonate-buffering characteristics of the raw water. PC2 was dominated by TOC and UV_{254}, with loadings of 0.91 and 0.84, respectively, and therefore primarily represents the organic matter load of the raw water. Thus, the major variations in water quality within Al-CD can be characterized by two principal dimensions, i.e., mineralization/carbonate chemistry and organic matter characteristics.

In contrast, the loadings of the water quality parameters in Fe-CD were more broadly distributed across the PCs. The first four PCs explained 81.6% of the total variance. In PC1, TOC, UV_{254}, alkalinity, and total hardness had loadings of 0.80, 0.73, 0.69, and 0.62, respectively, indicating that this component captures the coupled variation between organic indicators and mineralization/carbonate chemistry indicators. In PC2, alkalinity and total hardness had positive loadings of 0.63 and 0.70, respectively, whereas TOC and UV_{254} showed negative loadings of −0.51 and −0.59, respectively, indicating contrasting variation between these two groups of water quality characteristics.

Overall, Al-CD exhibited a relatively clear differentiation between major water quality characteristics, whereas Fe-CD showed more complex coupling among different water quality parameters. This difference may be associated with variations in source-water characteristics, hydrological processes, and plant operating conditions between the two datasets. Given the distinct overall patterns of water quality variation between Al-CD and Fe-CD, subsequent model training was conducted separately for the two datasets to better characterize coagulant dosage requirements under different raw-water conditions.

3.2. Machine Learning Modeling and Prediction Performance

MLR showed very limited predictive performance for coagulant dosage. The R^2, RMSE, and MAE values of Al-MLR on the test sets were 0.323, 1.882 mg/L Al, and 1.218 mg/L Al, respectively, while those of Fe-MLR were 0.515, 4.528 mg/L Fe, and 3.196 mg/L Fe, respectively (Figure S7). For Al-based coagulant dosage prediction, the optimized CatBoost, NN, RF, and SVR models achieved R^2 values of 0.76 ± 0.01, 0.76 ± 0.02, 0.69 ± 0.01, and 0.670, respectively, on the test set, with RMSE values of 1.1 ± 0.02, 1.15 ± 0.05, 1.28 ± 0.02, and 1.31 mg/L and MAE values of 0.65 ± 0.01, 0.73 ± 0.02, 0.74 ± 0.01, and 0.85 mg/L, respectively (Figures 3a and S7a–d). For Fe-CD, the optimized CatBoost, NN, RF, and SVR models achieved test-set R^2 values of 0.87 ± 0.01, 0.87 ± 0.01, 0.86 ± 0.03, and 0.82, respectively, with RMSE values of 2.31 ± 0.06, 2.31 ± 0.10, 2.45 ± 0.28, and 2.77 mg/L and MAE values of 1.53 ± 0.03, 1.54 ± 0.04, 1.55 ± 0.10, and 1.84 mg/L, respectively (Figures 3b and S8e–h). CatBoost and NN showed comparable predictive performance and generally outperformed RF, whereas SVR exhibited the lowest predictive performance among the four models. Given the right-skewed distribution of Fe dosages, a sensitivity analysis was conducted by excluding the highest-dose observations to assess their influence on model performance. The results showed that RMSE and MAE changed only modestly, suggesting that these observations had a comparatively limited influence on the overall absolute prediction errors. In contrast, R^2 was sensitive to their inclusion (Table S8).

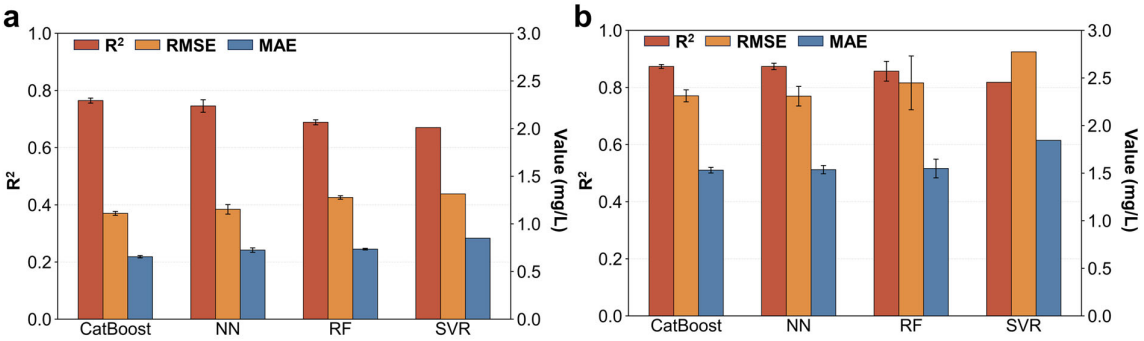


Figure 3. Variability in model predictive performance (R^2, RMSE, and MAE) across 21 runs using different random seeds (42–62): (**a**) Al-based coagulant dosage prediction; (**b**) Fe- based coagulant dosage prediction. Each run used an 80:20 training–test split. Detailed results are provided in Table S9.

The higher predictive accuracy of NN may be related to the complex nonlinear and coupled relationships among water quality parameters, which can be effectively captured through its multilayer nonlinear transformations. Although both CatBoost and RF are tree-based ensemble methods, CatBoost employs a boosting strategy that iteratively learns from the residuals of preceding models, which may enable it to better capture the complex nonlinear relationships and feature interactions between water quality parameters and coagulant dosage. This may explain its better predictive performance than RF. In contrast, the relatively lower predictive performance of SVR may be associated with its sensitivity to the choice of kernel function and hyperparameters, as well as the difficulty of fully capturing the complex nonlinear relationships and feature interactions present in the current dataset. The comparable predictive accuracy of CatBoost and NN, despite their fundamentally different learning mechanisms, may suggest that the information available for ML is approaching the upper limit under the current set of input features. The models explained approximately 77% of the variance in coagulant dosage, while the remaining approximately 23% was not captured by the models and may be attributable to unmeasured process variables and raw-water characteristics, such as flocculation mixing intensity, hydraulic conditions during rapid mixing and flocculation, and variations in the molecular composition of natural organic matter. Notably, SVR showed relatively lower predictive performance in both datasets, consistent with previous research based on one year of monitoring data from a single water treatment plant for coagulant dosage prediction [8]. This consistency suggests that SVR may have a relatively limited ability to capture the complex nonlinear relationships between water quality parameters and coagulant dosage.

To further assess the generalizability of the CatBoost models to previously unseen WTPs, an additional evaluation was conducted using a WTP-grouped data split, in which data from each WTP were assigned exclusively to either the training or test set. For Al-based coagulant dosage prediction, the training set included 1233 records from 125 WTPs, while the test set included 358 records from 14 WTPs. For Fe-based coagulant dosage prediction, the training set included 376 records from 38 WTPs, while the test set included 111 records from 4 WTPs. The models were trained using the optimal hyperparameter combinations identified under the original random-split scheme. Under the WTP-grouped split, the CatBoost models achieved R^2, RMSE, and MAE values of 0.68, 1.29 mg/L Al, and 0.77 mg/L Al, respectively, for Al-based coagulant dosage prediction, and 0.69, 2.21 mg/L Fe, and 1.64 mg/L Fe, respectively, for Fe-based coagulant dosage prediction. The lower R^2 values under the WTP-grouped split suggest that random splitting may overestimate the models' generalizability to unseen WTPs. However, the changes in

RMSE and MAE were relatively modest compared with the changes in R^2, suggesting that the choice of data-splitting strategy had a more pronounced effect on R^2 than on the absolute prediction errors.

3.3. Model Interpretation and Feature Importance Analysis

To elucidate the contributions of different water quality and operational parameters to coagulant dosage predictions, SHAP values and feature interactions were analyzed for the trained Al-CatBoost and Fe-CatBoost models. Given the comparable predictive performance of CatBoost and NN for Al-CD and the relatively small performance difference between the two models for Fe-CD, as well as the fact that TreeExplainer can provide exact SHAP values for supported tree models under specified assumptions regarding feature dependence, whereas DeepExplainer estimates SHAP values for neural networks, CatBoost was selected for SHAP analysis **Error! Reference source not found.**.

For Al-based coagulant dosage prediction, TOC, total hardness, UV_{254}, and alkalinity were the four most important features, with mean absolute SHAP values of 0.451, 0.330, 0.322, and 0.305 mg/L Al, corresponding to normalized importance values of 18.3%, 13.4%, 13.1%, and 12.4%, respectively (Figure 4, left). TOC and UV_{254} jointly accounted for 31.4% of the total importance, while total hardness and alkalinity accounted for an additional 25.8%, highlighting the importance of organic matter characteristics and mineralization/buffering-related water quality parameters in model predictions.

The contribution of TOC to the Al-CatBoost predictions exhibited a nonlinear pattern. Samples with relatively low TOC generally had negative SHAP values (Figure 4, right), whereas the contribution gradually approached zero with increasing TOC and shifted to predominantly positive values at approximately 4 mg/L C (Figure S10a), indicating that higher TOC was associated with higher predicted coagulant dosages in the current model. UV_{254} showed a similar overall transition from negative to positive contributions with increasing feature values (Figure S10c). In contrast, total hardness exhibited a contribution pattern opposite to those of the organic matter-related parameters (Figure S10b).

Feature interactions were further examined (Figures S11 and S12). The total hardness–alkalinity pair showed the highest mean absolute SHAP interaction value (0.191 mg/L Al), while other feature interactions were generally weak. However, the interaction contributions were widely dispersed across samples without a clear trend (Figures S13 and S14), suggesting that the observed joint contribution does not provide sufficient evidence for a consistent synergistic relationship between total hardness and alkalinity.

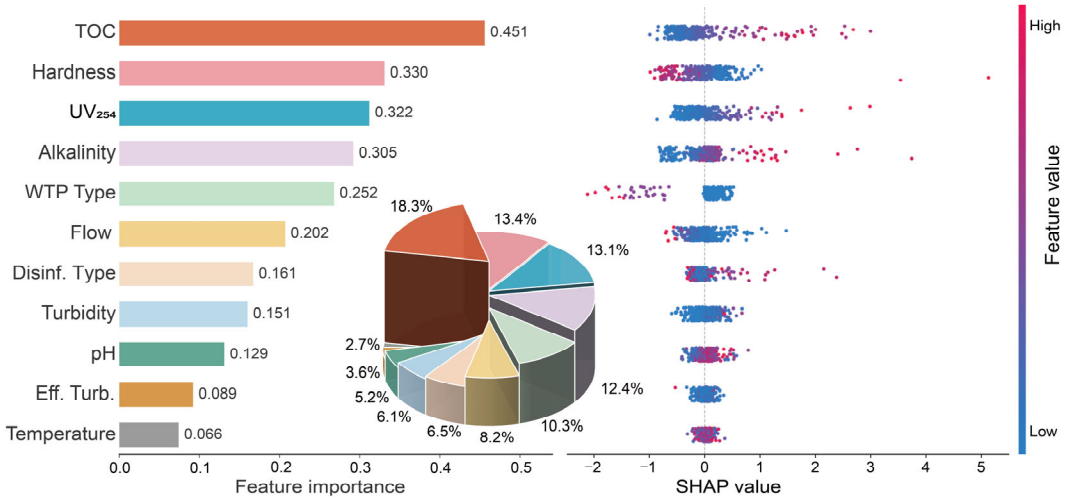


Figure 4. SHAP analysis of the Al-CatBoost model. (**Left**), mean absolute SHAP values; middle, relative feature importance; (**Right**), SHAP beeswarm plot.

Similar to the Al-CatBoost model, TOC, alkalinity, UV_{254}, and total hardness were the four most important features in Fe-CatBoost (Figure 5). The SHAP contributions of TOC and UV_{254} transitioned from negative to positive with increasing feature values, whereas those of alkalinity and total hardness showed the opposite trend. Thus, higher organic matter levels were associated with higher predicted Fe coagulant dosages, while higher alkalinity and hardness were associated with lower predicted dosages. Feature interactions were generally weak, indicating that Fe-CatBoost predictions were driven primarily by the individual contributions of the water quality parameters.

TOC was the most important feature in both the Al- and Fe-CatBoost models, with its SHAP contributions generally transitioning from negative to positive as TOC increased. This consistent pattern highlights the importance of organic matter in predicting coagulant dosage, consistent with the well-established role of natural organic matter (NOM) in coagulation. NOM can be removed through interactions with flocs via charge neutralization, adsorption, and sweep. Accordingly, higher organic matter levels may generally be associated with greater coagulant requirements.

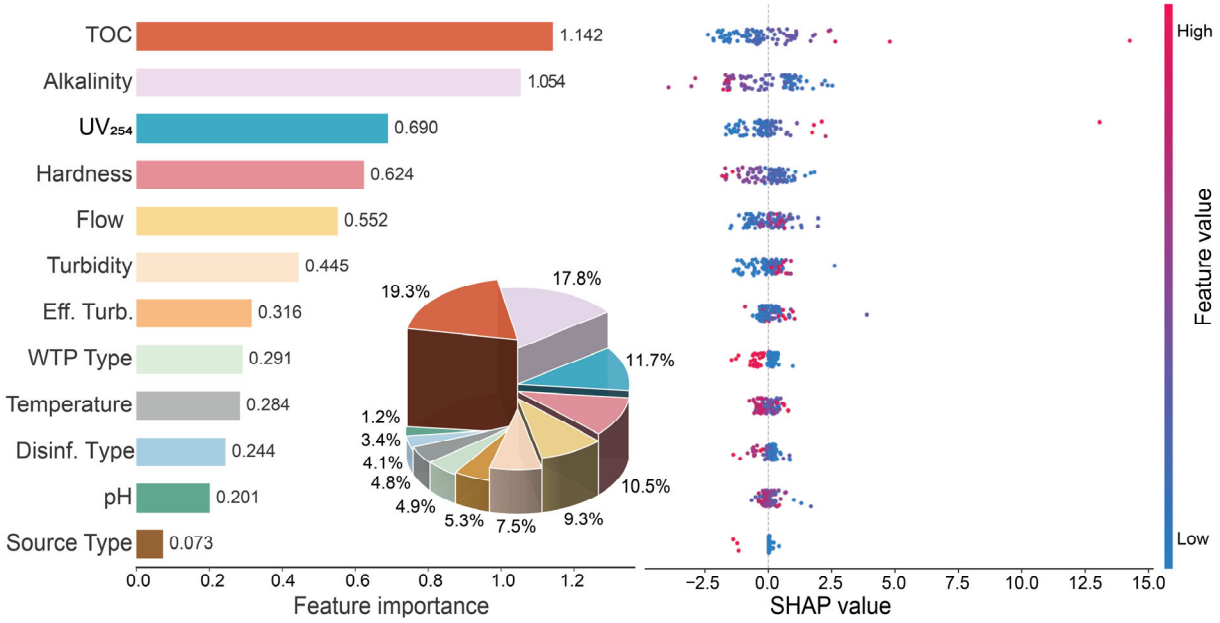


Figure 5. SHAP analysis of the Fe-CatBoost model. (**Left**), mean absolute SHAP values; middle, relative feature importance; (**Right**), SHAP beeswarm plot.

SHAP analysis has increasingly been used to interpret coagulant dosage prediction models, but the key features identified across studies are not entirely consistent. For example, Feng et al. used influent flow rate, temperature, pH, NH_3-N, turbidity, conductivity, COD_{mn}, effluent pH, and effluent turbidity as input features and identified influent conductivity, influent NH_3-N, effluent pH, COD_{mn}, and temperature as the five most important features based on a RF model (RMSE = 0.89 and R^2 = 0.96) [23]. Wang et al. developed XGBoost models for three water treatment plants using influent flow rate, temperature, turbidity, pH, effluent pH, effluent turbidity, and hypochlorous acid dosage as input features. Their models achieved RMSE values of 4.191–5.051 mg/L and R^2 values of 0.626–0.707, with influent temperature and turbidity identified as the primary influencing factors [24]. The differences are related to variations in the datasets and input

features. In particular, owing to limitations in the monitoring data, Wang et al. did not include organic matter indicators such as TOC and UV_{254} in their models, precluding further assessment of their contributions to coagulant dosage prediction. It should be noted that SHAP-based interpretations are inherently constrained by the feature set included in a model. They reflect the contributions of individual features to model predictions under a specific dataset and modeling framework, rather than establishing causal relationships or providing a universal ranking of factors governing coagulation. Therefore, comparisons of key influencing factors identified by SHAP across different studies should consider the data sources, plant operating conditions, and model input features, rather than relying directly on differences in SHAP importance rankings.

3.4. Model Application and Preliminary Performance Testing

The performance of the developed models was preliminarily assessed through jar tests using raw water collected from a WTP and additional data from published studies involving 12 WTPs in the United States and China [25,26]. In the jar tests, six independent water samples were randomly collected and treated with Al coagulant at 2.5 mg/L Al (n = 3) or 3.0 mg/L Al (n = 3). The basic water quality characteristics and model predictions are summarized in Table S10. The CatBoost model predictions yielded an MAE of 0.13 mg/L Al and an RMSE of 0.15 mg/L Al. The absolute percentage errors (APEs) ranged from 0.44% to 11.06%, with a mean absolute percentage error (MAPE) of 5.16%. For the data obtained from published studies (Table S10), the APEs ranged from 0.5% to 100.3%, with a MAPE of 21.09%.

The ICR database used in this study contains operational data from U.S. WTPs collected in 1997–1998. Despite the age of these data, the relationships between raw water quality and coagulant demand are governed by fundamental physicochemical principles, and parameters such as pH, turbidity, organic matter content, and alkalinity remain routinely measured and readily available in WTPs. In addition, the ICR and contemporary WTP datasets showed substantial overlap in the distributions of selected water quality parameters (Figure S15), providing a basis for considering the potential application of the developed models to contemporary WTPs.

To facilitate practical application, an online coagulant dosage prediction platform was developed integrating the CatBoost and NN models (icr-coagulant.streamlit.app). The platform integrates Al- and Fe-based coagulant dosage prediction with SHAP-based interpretation, providing the five most important contributing features and their directions of contribution (Figure S16). It should be noted that the prediction models were developed using historical WTP operational data, in which effluent turbidity reflects the actual treatment performance under historical operating conditions. However, the current effluent turbidity is generally unavailable before a dosing decision is made. Therefore, when using the online platform, users can specify a target effluent turbidity based on drinking water quality standards and actual treatment requirements to facilitate coagulant dosage prediction.

4. Conclusions

This study used the ICR database covering multiple WTPs and different temporal scales to construct separate datasets for Al- and Fe-based coagulant dosage prediction and systematically compared the predictive performance of four machine learning models, including CatBoost, neural networks (NN), random forests (RF), and support vector regression (SVR). The results revealed distinct water quality variation structures between the Al and Fe datasets. In the Al dataset, mineralization/carbonate buffering characteristics and organic matter characteristics exhibited relatively independent patterns of variation, whereas the Fe dataset showed more complex coupling among

organic matter, mineralization, and carbonate-related characteristics. These differences support the use of separate prediction models to better characterize the dosage requirements of different coagulant types.

All four models provided reasonably good predictions of coagulant dosage. CatBoost and NN achieved comparable predictive performance for both Al and Fe coagulant dosages, followed by RF, whereas SVR showed the lowest predictive accuracy. SHAP analysis further identified TOC, UV_{254}, alkalinity, and total hardness as common important water quality features in the Al- and Fe-CatBoost models. TOC showed the greatest feature contribution in both models, with its SHAP values generally transitioning from negative to positive contributions as TOC increased. UV_{254} exhibited a similar pattern, whereas the effects of alkalinity and total hardness varied with coagulant type and feature value ranges. Overall, feature interactions were relatively weak, indicating that model predictions were driven primarily by the individual contributions of the input features.

Preliminary jar-test experiments and testing with data from 12 WTPs reported in published studies provided additional evidence of the Al-CatBoost model's predictive performance under plant-relevant conditions, with MAPE values of 5.16% and 21.09%, respectively. Building on these results, an online platform integrating coagulant dosage prediction and SHAP-based interpretation was developed to enable rapid prediction of Al- and Fe-based coagulant dosages and identification of the key contributing water quality features. Overall, the developed framework provides a data-driven and interpretable approach to coagulant dosage estimation, with the potential to improve dosing responsiveness and reduce the need for preliminary tests and manual effort.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su182111263/s1, Text S1. Data Cleaning and Preprocessing; Text S2. Model Training and Implementation; Text S3. Water Quality Measurement Methods; Text S4. Benchmark Comparison with Linear and Simple Models; Figure S1. Category merging: PCA scatter plots (method 1 of 4); Figure S2. Category merging: Mahalanobis distance heatmaps (method 2 of 4); Figure S3. Category merging: PERMANOVA p-value heatmaps (method 3 of 4); Figure S4. Category merging: Ward hierarchical clustering dendrograms (method 4 of 4); Figure S5. PCA scree plots for AL and FE; Figure S6. Residual diagnostics for the best-performing models; Figure S7. Comparison of observed and predicted coagulant dosages using multiple linear regression (MLR) models: (a) Al-MLR and (b) Fe-MLR. The red dashed line represents the ideal 1:1 agreement line; Figure S8. Measured versus predicted dosages for CatBoost, RF, and SVR (AL and FE); Figure S9. Training curves for four machine learning models. Panels (a–d) show the AL models, ordered as CatBoost, NN, RF, and SVR. Panels (e–h) show the FE models in the same order. Red curves show training-set RMSE. Horizontal axes represent boosting iterations, epochs, number of trees, and training set size for CatBoost, NN, RF, and SVR, respectively. RMSE is expressed in mg/L Al or mg/L Fe, as appropriate; Figure S10. SHAP dependence plots for the main contributing features; Figure S11. SHAP main-effect and interaction contribution matrix for the Al model; Figure S12. SHAP main-effect and interaction contribution matrix for the Fe model; Figure S13. SHAP interaction between total hardness and alkalinity in the Al model; Figure S14. SHAP interaction between TOC and UV_{254} in the Fe model; Figure S15. Comparison of selected water quality parameters between the SYR4 dataset (2012–2019) and the ICR dataset (Al and Fe, 1997–1998): (a) pH, (b) TOC, (c) UV_{254}, and (d) total alkalinity. Diamonds and horizontal lines represent the means and medians, respectively; Figure S16. Screenshots of the intelligent dosing platform; Table S1. Descriptive statistics for all modeling variables; Table S2. Category distributions before and after merging; Table S3. Pairwise category similarity and merging recommendations; Table S4. CatBoost hyperparameters; Table S5. Neural network hyperparameters; Table S6. Random forest hyperparameters; Table S7. Support vector regression hyperparameters; Table S8. Sensitivity analysis of Fe coagulant dosage prediction performance after excluding the four highest-dose

observations from the test set; Table S9. Mean and standard deviation of model predictive performance (R2, RMSE, and MAE) across 21 runs using different random seeds (42–62); Table S10. Raw water quality characteristics and coagulant dosage data for the six samples used in jar tests and the 12 water treatment plants reported in published studies; Table S11. Sensitivity of CatBoost prediction performance to the exclusion of minority plant-type categories.

Author Contributions: Q.-K.L.: Investigation, Data curation, Visualization, Writing—original draft. Z.-J.F.: Data curation, Validation. Z.-Y.X.: Data curation, Validation. C.-M.X.: Experimentation, Data curation, Writing—original draft. Y.-T.H.: Data curation, Validation. X.-m.W., Y.X. and W.-J.F.: Conceptualization, Supervision, Writing—review and editing, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: The National Natural Science Foundation of China (Grant No. 22206028 and 42501576), the Foundation of Talent Projects of Guangxi (Grant No. AD23026071), the Guangxi Natural Science Foundation (grant no. 2025GXNSFBA069103), the Innovation Project of Guangxi Graduate Education (Grant No. XYCS2026096), the Jing-Jin-Ji Regional Integrated Environmental Improvement-National Science and Technology Major Project (Grant No. 2026ZD1213700).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original Disinfection Byproducts Information Collection Rule (DBP ICR) Aux 1 data analyzed in this study are publicly available from the U.S. Environmental Protection Agency at https://www.epa.gov/dwsixyearreview/supplemental-data-six-year-review-3, accessed on 4 September 2025.

Acknowledgments: We acknowledge the Guangxi Bagui Young Top Talent Program. We thank Hai-Yan Xiao for her valuable suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Matilainen, A.; Vepsäläinen, M.; Sillanpää, M. Natural organic matter removal by coagulation during drinking water treatment: A review. *Adv. Colloid Interface Sci.* **2010**, *159*, 189–197.
2. Franceschi, M.; Girou, A.; Carro-Diaz, A.M.; Maurette, M.T.; Puech-Costes, E. Optimisation of the coagulation-flocculation process of raw water by optimal design method. *Water Res.* **2002**, *36*, 3561–3572.
3. Bérubé, D.; Dorea, C. Optimizing alum coagulation for turbidity, organics, and residual Al reductions. *Water Sci. Technol. Water Supply* **2008**, *8*, 505–511.
4. Ratnaweera, H.; Fettig, J. State of the Art of Online Monitoring and Control of the Coagulation Process. *Water* **2015**, *7*, 6574–6597.
5. Yavich, A.A.; Van De Wege, J. Chemical feed control using coagulation computer models and a streaming current detector. *Water Sci. Technol.* **2013**, *67*, 2814–2821.
6. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J.; et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **2020**, *171*, 115454.
7. Lin, S.; Kim, J.; Hua, C.; Park, M.-H.; Kang, S. Coagulant dosage determination using deep learning-based graph attention multivariate time series forecasting model. *Water Res.* **2023**, *232*, 119665.
8. Zhang, K.; Achari, G.; Li, H.; Zargar, A.; Sadiq, R. Machine learning approaches to predict coagulant dosage in water treatment plants. *Int. J. Syst. Assur. Eng. Manag.* **2013**, *4*, 205–214.
9. Kim, C.M.; Parnichkun, M. Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system. *Appl. Water Sci.* **2017**, *7*, 3885–3902.
10. Wu, G.-D.; Lo, S.-L. Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network. *Expert Syst. Appl.* **2010**, *37*, 4974–4983.

11. Heddam, S.; Bermad, A.; Dechemi, N. ANFIS-based modelling for coagulant dosage in drinking water treatment plant: A case study. *Environ. Monit. Assess.* **2012**, *184*, 1953–1971.
12. Kim, J.; Hua, C.; Kim, K.; Lin, S.; Oh, G.; Park, M.-H.; Kang, S. Optimizing coagulant dosage using deep learning models with large-scale data. *Chemosphere* **2024**, *350*, 140989.
13. Xu, S.; Wang, S.; Yang, Y.; Zheng, L.; Xiao, F. Importance analysis of key parameters for coagulation dosage prediction based on machine learning models. *Process Saf. Environ. Prot.* **2025**, *205*, 108152.
14. Wang, D.; Chen, L.; Li, T.; Chang, X.; Ma, K.; You, W.; Tan, C. Successful prediction for coagulant dosage and effluent turbidity of a coagulation process in a drinking water treatment plant based on the Elman neural network and random forest models. *Environ. Sci. Water Res. Technol.* **2023**, *9*, 2263–2274.
15. Archer, A.D.; Singer, P.C. An evaluation of the relationship between SUVA and NOM coagulation using the ICR database. *J.-Am. Water Work. Assoc.* **2006**, *98*, 110–123.
16. United States Environmental Protection Agency. Supplemental Data for Six-Year Review 3. Available online: https://www.epa.gov/dwsixyearreview/supplemental-data-six-year-review-3 (accessed on 4 November 2025).
17. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018; Volume 32.
18. Dietterich, T.G. Ensemble methods in machine learning. In *Proceedings of the International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000.
19. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
20. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **1997**, *28*, 779–784
21. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
22. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.
23. Feng, L.; Zhang, Y.; Wei, X.; Wang, M.; Niu, Z.; Wang, C. Interpretable prediction of coagulant dosage in drinking water treatment plant based on automated machine learning and SHAP method. *J. Water Process Eng.* **2025**, *75*, 107925.
24. Wang, P.; Zheng, X.; Xu, S.; Jiang, C.; Huang, X.; An, G.; Lin, J.; Xie, Y.; Shi, B.; Zhuang, X. Interpretable machine learning for coagulant dosage optimization in drinking water treatment plants: A comparative study of distinct process configurations. *Sep. Purif. Technol.* **2026**, *397*, 138052.
25. Ma, J.; Liu, W. Effectiveness of ferrate (VI) preoxidation in enhancing the coagulation of surface waters. *Water Res.* **2002**, *36*, 4959–4962.
26. Alansari, A.Y. A Comprehensive Study of Drinking Water Coagulation with Aluminum Sulfate. Ph.D. Thesis, The University of North Carolina at Charlotte, Charlotte, NC, USA, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.