

Building a reliable coagulant dosage prediction framework for water purification processes using CatBoost and explainable AI (XAI)

Khaled Merabet, Il-Moon Chung, Amel Fouchal, Francesco Granata, Ozgur Kisi, Sungwon Kim, Salim Heddad & Mohammad Zounemat-Kermani

To cite this article: Khaled Merabet, Il-Moon Chung, Amel Fouchal, Francesco Granata, Ozgur Kisi, Sungwon Kim, Salim Heddad & Mohammad Zounemat-Kermani (2026) Building a reliable coagulant dosage prediction framework for water purification processes using CatBoost and explainable AI (XAI), Engineering Applications of Computational Fluid Mechanics, 20:1, 2706865, DOI: [10.1080/19942060.2026.2706865](https://doi.org/10.1080/19942060.2026.2706865)

To link to this article: <https://doi.org/10.1080/19942060.2026.2706865>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 19 Aug 2026.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

Building a reliable coagulant dosage prediction framework for water purification processes using CatBoost and explainable AI (XAI)

Khaled Merabet ^a, Il-Moon Chung ^b, Amel Fouchal ^c, Francesco Granata ^{d,e}, Ozgur Kisi ^{f,g,h}, Sungwon Kim ⁱ, Salim Heddam ^a and Mohammad Zounemat-Kermani ^j

^aFaculty of Science, Agronomy Department, Hydraulics Division, University 20 Août 1955 Skikda, Skikda, Algeria; ^bDepartment of Hydro Science and Engineering, Korea Institute of Civil Engineering and Building Technology, Goyang-si, Republic of Korea; ^cLEGHYD Laboratory, Department of Civil Engineering, University of Science and Technology Houari Boumediene, Algiers, Algeria; ^dDepartment of Civil and Mechanical Engineering (DICEM), University of Cassino and Southern Lazio, Cassino, Italy; ^eEuropean University of Technology, European Union, Cassino, Italy; ^fDepartment of Civil Engineering, Lübeck University of Applied Sciences, Lübeck, Germany; ^gDepartment of Civil Engineering, School of Technology, Ilia State University, Tbilisi, Georgia; ^hSchool of Civil, Environmental and Architectural Engineering, Korea University, Seoul, South Korea; ⁱDepartment of Railroad Construction and Safety Engineering, Dongyang University, Yeongju, Republic of Korea; ^jDepartment of Civil Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ABSTRACT

Coagulant dosage modelling has been widely investigated in the literature; however, existing methods often operate as black boxes, failing to provide clear practical explanations for decision-makers. To address this limitation, this study introduces an interpretable framework using SHAP and LIME to highlight the underlying practical problems. Furthermore, this is the first study of its kind in Algeria, bridging a critical regional gap in the literature. Yet, we investigate the application of boosting machine learning models to predict coagulant dose at the Taksebt drinking water treatment plant. Herein we compare between: (i) multiple linear regression (MLR), (ii) random forest regression (RFR), (iii) adaptive boosting (AdaBoost), (iv) categorical boosting (CatBoost), (v) gradient boosted regression trees (GBRT), (vi) Histogram Gradient Boosting (HistGBRT), (vii) Light Gradient Boosting Machine (LightGBM), (viii) Natural Gradient Boosting (NGBoost), and (ix) eXtreme Gradient Boosting (XGBoost). All models were trained using daily measured water quality variables, while aluminium sulphate ($\text{Al}_2(\text{SO}_4)_3 \cdot 18\text{H}_2\text{O}$) was the coagulant dose. The predictive performance was evaluated using root-mean-square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe efficiency (NSE), and the correlation coefficient (R). Global and local model interpretability were conducted using SHAP and LIME, thereby enabling feature importance ranking. To determine whether the observed differences among models were statistically meaningful, two complementary statistical tests were applied: the Diebold-Mariano test and the Kruskal-Wallis test. The results indicate that CatBoost delivered the most accurate predictions, achieving R, NSE, RMSE, and MAE values of 0.922, 0.849, 2.498, and 1.702 mg/L, respectively, and significantly outperforming the MLR model, which yielded an R of 0.713, a NSE value of 0.507, RMSE of 4.510 mg/L, and MAE of 3.576 mg/L, respectively. Based on SHAP analysis, UV_{254} was identified as the most influential feature, contributing 24% to the model predictions, whereas COU showed a negligible effect, accounting for only 5%.

Highlights

- Boosting models for coagulant dosage prediction.
- CatBoost provided the best numerical performances.
- UV_{254} was the most influential raw water quality variable.
- SHAP and LIME for model interpretability.

ARTICLE HISTORY

Received 13 May 2026
Accepted 15 July 2026

KEYWORDS

Coagulant dose; raw water quality variables; modelling; boosting models; SHAP; LIME

CONTACT Khaled Merabet  kh.merabet@univ-skikda.dz  Faculty of Science, Agronomy Department, Hydraulics Division, University 20 Août 1955 Skikda, Route El Hadaik, BP 26, Skikda 21000, Algeria; Il-Moon Chung  imchung@kict.re.kr  Department of Hydro Science and Engineering, Korea Institute of Civil Engineering and Building Technology, Goyang-si, 10223, Republic of Korea; Sungwon Kim  swkim1968@dyu.ac.kr  Department of Railroad Construction and Safety Engineering, Dongyang University, Yeongju 36040, Republic of Korea

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19942060.2026.2706865>.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

1. Introduction

The reliable provision of safe drinking water represents one of the most fundamental technological and societal imperatives of the modern era, underpinning public health, economic stability, and sustainable development. In conventional drinking water treatment plants (DWTPs), coagulation is a pivotal physicochemical process that destabilizes colloidal suspensions and promotes their aggregation into settleable flocs, thereby enabling the efficient removal of turbidity, organic matter, and microbial contaminants (K. Zhang et al., 2013). The effectiveness of this process exerts a direct and cascading influence on downstream treatment performance, affecting sedimentation efficiency, filtration reliability, chemical consumption, sludge production, and operational costs (Kim et al., 2024; S. Xu et al., 2026). Consequently, the precise determination of coagulant dosage represents not merely an operational parameter but a central determinant of treatment reliability, economic efficiency, and environmental sustainability.

Despite its operational importance, optimal coagulant dosing remains inherently difficult to determine due to the nonlinear, multivariate, and time-dependent interactions among numerous physicochemical variables, including turbidity, pH, temperature, conductivity, dissolved oxygen, ultraviolet absorbance, and organic matter concentration (Lin et al., 2023a; S. Xu et al., 2026). These relationships are further complicated by hydrological variability, seasonal fluctuations, and episodic environmental disturbances, which induce rapid and often unpredictable changes in raw water quality (Kim et al., 2024). Traditionally, coagulant dosage has been determined using empirical approaches, most notably jar tests and operator expertise. While widely adopted, these methods suffer from intrinsic limitations, including delayed response times, labour intensity, and reliance on subjective judgment, rendering them fundamentally inadequate for real-time process optimization under dynamically evolving conditions (Bressane et al., 2023).

To position the present contribution within the broader methodological landscape, the evolution of data-driven approaches to water-quality and coagulant-dosage prediction may be conceptualized as four progressive stages. In the first stage, traditional machine-learning methods such as multiple linear regression, support vector machines, and k-nearest neighbours established the feasibility of data-driven dosage estimation. In the second stage, ensemble-learning techniques, encompassing bagging-based architectures such as random forests and boosting-based architectures such as Gradient Boosting Regression Trees, XGBoost, LightGBM, and CatBoost, were introduced to capture the nonlinear and

multivariate interactions that single learners could not adequately represent. In the third stage, deep-learning architectures, including convolutional, recurrent, graph-attention, and transformer-based models, were adopted to extract hierarchical spatiotemporal features from high-resolution operational datasets. In the fourth and most recent stage, explainable artificial intelligence (XAI) has been integrated so that model behaviour may be rendered transparent and operationally trustworthy. The literature reviewed below is organized according to this four-stage perspective.

In accordance with this perspective, prior studies may be organized hierarchically into three methodological categories. The first category comprises linear and statistically grounded models, most notably multiple linear regression, which afford direct interpretability but are intrinsically limited in capturing the nonlinear, threshold-dependent interactions that govern coagulant demand (Etemadi & Khashei, 2021; Shams et al., 2021). The second category encompasses tree-based ensemble models, spanning bagging architectures such as random forests and boosting architectures such as Gradient Boosting Regression Trees, XGBoost, LightGBM, and CatBoost. Robustness within this category differs appreciably among algorithms: boosting learners that sequentially correct residuals tend to generalize more reliably under noisy and heterogeneous time-series conditions than single learners or bagging ensembles, while the integration of multi-source data fusion and signal-decomposition strategies has been shown further to enhance ensemble robustness in water-quality forecasting (M. An et al., 2026; S. Xu et al., 2026). The third category concerns XAI-based water-quality prediction, in which post-hoc attribution methods such as SHAP and LIME are coupled with predictive models in order to quantify feature contributions. Although global attribution has been widely reported, comparatively few investigations have addressed local, instance-level interpretation of time-series water-quality predictions; this gap has recently been addressed through the coupling of interpretable frameworks with transformer- and Kolmogorov-Arnold-network-based architectures and through advanced local-explanation strategies (Ding et al., 2026; Gharehbaghi et al., 2026a). The present study is situated within, and seeks to integrate, all three categories.

The inherent limitations of empirical approaches have motivated the development of predictive models that estimate coagulant dosage directly from operational water-quality data. Early machine learning applications demonstrated the feasibility of modelling coagulation processes using data-driven methods, with support vector machines and k-nearest neighbours achieving reliable

predictive performance across diverse treatment plants, thereby establishing the viability of machine learning for coagulant dosage prediction (K. Zhang et al., 2013). Subsequent studies significantly expanded this paradigm. For example, hybrid machine learning frameworks integrating decision tree models with optimization algorithms achieved correlation coefficients exceeding 0.96 and improved predictive accuracy by more than 70% compared to conventional empirical methods, demonstrating the ability of machine learning to capture complex nonlinear dependencies among process variables (Achite et al., 2024). Similarly, fuzzy inference system-based approaches demonstrated substantial reductions in prediction error, outperforming neural networks, random forests, and other conventional models by more effectively capturing uncertainty and nonlinear interactions within treatment systems (Bressane et al., 2023).

More recent developments have leveraged ensemble learning and deep learning techniques to further enhance predictive performance. Ensemble machine learning models, including random forests and hybrid boosting frameworks, have demonstrated strong predictive performance, reducing mean absolute percentage errors to below 5% while simultaneously improving treatment efficiency and reducing chemical consumption (S. Xu et al., 2026). In parallel, hybrid machine learning approaches integrating metaheuristic optimization techniques have achieved predictive accuracies exceeding 95%, highlighting the capacity of ensemble-based frameworks to model complex treatment dynamics (Kalhor-mohammadi & Khoramipour, 2025; Randive et al., 2025). These advances underscore the growing role of ensemble learning as a robust and effective tool for modelling nonlinear environmental systems.

Deep learning approaches have further advanced the state of the art by enabling the extraction of hierarchical temporal and nonlinear features from high-resolution operational datasets. Graph attention-based deep learning models have demonstrated coefficients of determination exceeding 0.94, significantly outperforming traditional machine learning methods by capturing complex temporal dependencies and feature interactions (Lin et al., 2023a). Comparative analyses between conventional machine learning and deep learning frameworks have further confirmed the superior generalization capability of deep architectures, particularly under highly nonlinear and heterogeneous physicochemical conditions, highlighting their ability to capture complex interactions that remain unresolved by traditional ensemble and regression models (Lin et al., 2023b). Similarly, deep learning frameworks incorporating convolutional architectures have achieved Nash–Sutcliffe efficiency values exceeding 0.93 and correlation coefficients approaching

0.97, demonstrating their ability to capture periodicity and nonlinear dynamics inherent in water treatment processes (Kim et al., 2024). These improvements have translated into tangible operational benefits, including reductions in chemical consumption of more than 20% while maintaining compliance with treated water quality standards (Kim et al., 2024). Furthermore, explainable artificial intelligence approaches have enabled the identification of key physicochemical drivers governing coagulant demand, with variables such as conductivity, colour, and turbidity consistently identified as dominant predictors (Fouchal et al., 2025; Namane et al., 2025).

Despite these advances, important limitations remain. Deep learning models often require extensive training datasets, substantial computational resources, and careful hyperparameter tuning, which may constrain their operational deployment. Moreover, their inherent architectural complexity may limit interpretability, which remains a critical requirement for safety-critical infrastructure systems such as water treatment facilities. In this context, boosting algorithms represent a particularly promising yet comparatively underexplored modelling paradigm. By sequentially aggregating weak learners into a robust predictive ensemble, boosting methods can effectively approximate highly nonlinear functional relationships while maintaining strong generalization capability and computational efficiency (Friedman, 2001). Modern boosting implementations, including Gradient Boosting Regression Trees, XGBoost, LightGBM, and CatBoost, have demonstrated exceptional performance across diverse environmental and engineering applications, achieving predictive accuracies exceeding 95% while maintaining robustness to noise and heterogeneity in real-world datasets (Randive et al., 2025; S. Xu et al., 2026).

The need for robust predictive frameworks is particularly evident in water treatment systems characterized by pronounced physicochemical variability. The Taksebt water treatment plant in Algeria provides a compelling and representative case study in this regard. This facility treats surface water exhibiting substantial temporal variability in key physicochemical parameters, including turbidity, conductivity, temperature, dissolved oxygen, ultraviolet absorbance, and colour, all of which exert nonlinear and interacting influences on coagulant demand (Fouchal et al., 2025). Previous investigations conducted at this facility have demonstrated that advanced machine learning and deep learning models can effectively predict coagulant dosage while identifying conductivity and colour as dominant controlling variables, underscoring the importance of capturing nonlinear interactions among physicochemical parameters

(Fouchal et al., 2025). Despite these advances, a critical gap persists in developing predictive models that simultaneously achieve high predictive accuracy, computational efficiency, robustness, and operational interpretability. Recent research has predominantly focused on deep learning architectures, which have demonstrated remarkable predictive performance (Fouchal et al., 2025; Kim et al., 2024; Lin et al., 2023a). However, these models often require large training datasets, substantial computational resources, and complex hyperparameter tuning, which may limit their practical deployment in operational water treatment environments characterized by limited data availability and strict reliability requirements. Furthermore, their intrinsic architectural complexity may hinder interpretability, which remains essential for process control and operational decision-making.

Accordingly, the present study addresses this critical research gap by developing and systematically evaluating advanced boosting-based models for coagulant dosage prediction using real-world operational data from the Taksebt water treatment plant. Specifically, this study aims to achieve three primary objectives. First, it evaluates the predictive performance of state-of-the-art boosting algorithms including AdaBoost, Gradient Boosting Regression Trees, XGBoost, LightGBM, and CatBoost in modelling coagulant dosage under complex and variable physicochemical conditions. Second, it identifies and quantifies the relative importance of water quality parameters governing coagulant demand, thereby improving mechanistic understanding of coagulation dynamics. Third, it establishes a robust, interpretable, and operationally deployable predictive framework that supports intelligent, adaptive process control in real-world water treatment systems. By integrating advanced ensemble learning techniques with high-resolution operational data, this study advances the emerging paradigm of intelligent, data-driven water treatment and represents a significant step towards realizing autonomous, adaptive, and resource-efficient water treatment infrastructure.

The novelty of the present study resides not in the isolated application of individual boosting algorithms, which have indeed been compared in earlier publications, but in the integration of three methodological components into a single, coherent, and operationally oriented framework. First, a hierarchical feature-elimination experiment is conducted, in which six progressively reduced input configurations (from seven predictors down to two) are applied uniformly across all nine algorithms, so that algorithmic robustness may be isolated from data-related effects and the minimal predictor set required for dependable prediction may be identified. Second, the statistical significance of the observed

performance differences is established by means of two complementary tests: the Diebold-Mariano test of equal predictive accuracy and the non-parametric Kruskal-Wallis's test of distributional equivalence, rather than being inferred from point-estimate metrics alone. Third, model interpretability is examined through the combined application of global and local XAI techniques, whereby SHAP and LIME are jointly employed in order to reconcile global feature-importance rankings with instance-level explanations. The simultaneous integration of these three components distinguishes the present contribution from earlier horizontal comparisons of boosting models and provides a transferable methodology for transparent, data-driven coagulant-dosage control.

2. Methods

2.1. Multiple linear regression (MLR)

Multiple linear regression (MLR) is a statistical method used to model the linear relationship between a single criterion and two (or more) predictor variables. It expands plain linear regression by combining numerous predictor variables, allowing analysts to predict effects based on a combination of numerous components (Agirre-Basurko et al., 2006; Nimon & Oswald, 2013). It has a powerful advantage for solving predictive problems in that it allows analysts to intuitively see in which direction (i.e. positive or negative) and to what extent a specific variable affects the criterion variable (Y) through the regression coefficient attached to each predictor variable (X). On the other hand, it basically assumes that the relationship between predictor and criterion variables is *linear* (*straight line*), and requires the variables to be independent of each other, so it has the disadvantage of having difficulty capturing complex nonlinear relationships and interactions between variables in the real world (Etemadi & Khashei, 2021; Shams et al., 2021).

The core formula of MLR is to define the predicted value of the criterion variable as a linear function of several predictor variables, and to estimate the coefficients (i.e. weights) that best supply the relationship between the predictor variables. That is, it is a process of finding an optimal line by minimizing the sum of squared errors using the least squares method, which is the sum of the squared distances between the actual and predicted values in the dataset. Therefore, the mathematical relation in MLR can be provided as formula (1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon \quad (1)$$

where Y = criterion variable; $X_1, X_2, \dots, X_k$ = predictor variables; β_0 = intercept; $\beta_1, \beta_2, \dots, \beta_k$ = regression coefficients; and ϵ = residuals, respectively. To ensure

predictive results with MLR, several key assumptions must hold, including the absence of multicollinearity, independence, homoscedasticity, normality, and linearity (Guillén-Casla et al., 2011). The analysts can collect more historic documents and reports, including coagulant dosage prediction, utilizing MLR to assist them in understanding clearly from Agirre-Basurko et al. (2006), Chakraborty and Goswami (2017), and Achite et al. (2024).

2.2. Random forest regression (RFR)

Random forest regression (RFR) operates within a composite learning framework, building a collection of classification trees over the course of the learning process (Breiman, 2001). This collective method leverages the combined knowledge collected from numerous tree architectures to improve predictive accuracy (Achite et al., 2023; Z. Zhou et al., 2023). It successfully manages sophisticated nonlinear interdependencies and complex dataset architecture, and also presents excellent computational achievement when handling massive and multi-dimensional datasets (Ahmad et al., 2018; J. Zhang et al., 2025). The basic component of RFR is a decision tree architecture. However, unlike individual decision tree implementation, it combines additional probabilistic elements throughout the tree-building process, including feature selection and split point identification procedure (Breiman, 2001; L. Feng et al., 2025), and excels at managing complex dataset interdependencies and preventing over-specialization of the model. Also, variations generated by the probabilistic tree development process ensure effective model mobility to new datasets, thereby improving overall stability (S. Xu et al., 2026; Z. Zhou et al., 2023).

RFR employs bootstrap aggregation to generate individual trees from the resampled subgroups of the original training dataset. The addressed strategy enhances model flexibility and reduces predictive variability by avoiding dependence on sources of a specific dataset (Z. Zhou et al., 2023). The mathematical formulation of RFR can be expressed as the following formula (2).

$$\hat{f}_{rf}^K(X) = \frac{1}{K} \sum_{k=1}^K T_k(X) \quad (2)$$

where, $\hat{f}_{rf}^K(X)$ = RFR prediction utilizing K trees for input X ; $T_k(X)$ = predictive performance for the k -th decision tree; and $\frac{1}{K} \sum_{k=1}^K T_k(X)$ = the average of predictive performances across all trees, respectively.

The greatest strength of RFR is that it is robust to data noise because it utilizes a bagging method to independently train multiple decision trees and then averages the

results. Therefore, it effectively prevents overfitting, in which the model is overly fitted to the training data. On the contrary, the main drawback of RFR is that it is an ensemble of decision trees, and since decision trees output the average value of leaf nodes to which the input data belongs, they cannot predict values that are outside the range of training data (S. Xu et al., 2026; Z. Zhou et al., 2023). For comprehensive knowledge on the development and application of RFR, including coagulant dosage prediction, see the significant research articles by Ahmad et al. (2018), L. Feng et al. (2025), and S. Xu et al. (2026).

2.3. Adaptive boosting (AdaBoost)

Adaptive boosting (AdaBoost) is a sequential ensemble method that builds multiple weak learners by leveraging different training subsets randomly selected from an initial training dataset (Karupiah et al., 2026). During each training iteration, specific weights are assigned, which serve two purposes: first, they lead to the learning procedure for individual hypotheses, and second, they assist the error calculation for each hypothesis across the dataset, while also indicating the relative importance of each sample dataset (Gharehbaghi et al., 2026b; S. Wang et al., 2024). It includes the superior performance when solving complex classification problems and handling high-dimensional datasets with multiple essences. The extent to which each predictor variable contributes to the final outcome is correlated with the accuracy achieved when evaluated using a weighted subset of training samples (D. C. Feng et al., 2020; L. Wang et al., 2022).

In addition, it displays resistance to overfitting throughout the learning process, a feature that has been theoretically validated and empirically supported in previous research (Shi et al., 2024). Therefore, weight adjustments for individual cases should be made according to the method in formula (3) (S. Wang et al., 2024).

$$W_i^{t+1} = w_i^t \exp(\alpha_t L_i^t) \quad (3)$$

where w_i^t = weight of the i -th training sample at iteration t ; W_i^{t+1} = updated weight of the same sample for the next boosting procedure; α_t = coefficient assigned to the weak regressor at iteration t ; and L_i^t = loss/error for the i -th data under the weak regressor, respectively. The previously calculated weight values must be normalized so that they sum to 1.0 utilizing formula (4).

$$W_i^{t+1} = \frac{W_i^{0(t+1)}}{\sum_{i=1}^n W_i^{0(t+1)}} \quad (4)$$

where, $W_i^{0(t+1)}$ = the unnormalized weight of sample i at iteration $t+1$; $\sum_{i=1}^n W_i^{0(t+1)}$ = the sum of all

unnormalized weights across the n training samples; and W_i^{t+1} = the final normalized weight of sample i at iteration $t+1$, respectively.

AdaBoost has the advantage of creating a learner with very strong performance by linking simple and weak learners together for solving predictive problems because it utilizes a sequential learning method that assigns higher weights to data with errors from previous learners, allowing the next learner to focus on learning them. However, it has a tendency to overweight incorrect problems, so it tends to mistake incorrect noise or unusual outliers as important data that must be corrected and stick to them until the end. Therefore, if there are outliers or noise in the data, there is a high risk of the model's performance deteriorating rapidly or overfitting (Shi et al., 2024; L. Wang et al., 2022). Conventional application of AdaBoost to solve nonlinear prediction and computational problems, including coagulant dosage prediction, can be found in previous articles, such as D. C. Feng et al. (2020), S. Wang et al. (2024), and Gharehbaghi et al. (2026b).

2.4. Categorical boosting (CatBoost)

Categorical boosting (CatBoost) is a boosting-based predictive framework that operates via a recurrent tree-based learning mechanism. A core feature that sets CatBoost apart is its ability to handle non-numeric feature attributes during training (Prokhorenkova et al., 2018; Randive et al., 2025). These features enable CatBoost to efficiently handle non-numeric variables without extensive data preparation, including binary encoding conversion. It successfully identifies complex relationships and interdependencies in the dataset by leveraging natural organizational patterns within categorical attributes, resulting in accurate prediction (Dasi et al., 2023; Y. Zhang et al., 2020). CatBoost uses an optimization-based algorithm (e.g. a greedy approach) to generate successive approximations across computation cycles, aiming to minimize a predefined objective function. It also leverages a robust method known as sequential target encoding to transform non-numerical features into quantitative representations, which can be formalized as shown in formula (5) (Ahmed et al., 2024; Prokhorenkova et al., 2018).

$$X_{\sigma_p,k} = \frac{\sum_{j=1}^{p-1} [X_{\sigma_j,k} = X_{\sigma_p,k}] Y_{\sigma_s} + \beta \cdot P}{\sum_{j=1}^{p-1} [X_{\sigma_j,k} = X_{\sigma_p,k}] + \beta} \quad (5)$$

where, $X_{\sigma_p,k}$ = value of feature k for partition σ_p ; Y_{σ_s} = output associated with sample K in partition σ ; β =

smoothing parameter (i.e. regularization to avoid overfitting when leaf counts are small), and P = some prior prediction, respectively.

CatBoost handles categorical data with great sophistication when solving predictive problems, without complex preprocessing such as one-hot encoding. Also, unlike conventional target encoding methods, which suffer from *target leakage* (i.e. learning by knowing the correct answers in advance), it accounts for data order and eliminates this problem at its source. In addition, the sequential boosting method, which randomly shuffles data multiple times and computes statistics at each step to prevent *target leakage*, significantly increases computational complexity. As a result, the learning process is very heavy memory-intensive (Prokhorenkova et al., 2018; Y. Zhang et al., 2020). For a more in-depth exploration of development and application of CatBoost, including coagulant dosage prediction, consider our comprehensive description in Prokhorenkova et al. (2018), Dasi et al. (2023), and Ahmed et al. (2024).

2.5. Gradient boosted regression trees (GBRT)

Gradient boosted regression trees (GBRT) are a combined predictive model based on the principles of the boosting algorithm. It utilizes multiple regression-based tree architectures as its fundamental model, and employs a forward stepwise method. At each step, a new weak classifier is developed along the gradient direction to minimize the residual, thereby reducing the cumulative predictive error from the weak classifiers in the previous step (Jiang et al., 2022; Nie et al., 2021; Zhu et al., 2024). At the same time, it modifies the weighting scheme applied to the training instances, allowing continuous learning to produce new baseline predictors that perfectly capture the data characteristics, and improve the predictive accuracy of weak models. Finally, the integrative approach combines the results of all weak models via linear integration, yielding predictions with higher accuracy (He et al., 2020; Jiang et al., 2022). In this method, the initialization of fundamental predictor variables proceeds according to the approach based on the following formula (6) (Liao et al., 2020).

$$f_0(x) = \arg \min \sum_{i=1}^n L(y_i, c) \quad (6)$$

where, $f_0(x)$ = the initial prediction function before any boosting iterations; y_i = the actual value of criterion variable for sample i ; and $L(y_i, c)$ = the loss function between the actual value y_i and a constant prediction c , respectively. In GBRT, the calculation of estimated residual values is also performed through a mathematical

approach as shown in the following formula (7) (Liao et al., 2020).

$$r_{mi} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{m-1}(x)} \quad (7)$$

where, r_{mi} = the pseudo-residual for sample i at iteration m ; $L(y_i, f(x_i))$ = the chosen loss function; and $f_{m-1}(x)$ = the current boosted model before adding the m -th tree, respectively.

GBRT has the advantage of maximizing performance for solving predictive problems by using a sequential method in which a new tree learns the residuals that the previous tree did not predict, and gradually improves on the mistakes of the previous stage in the next stage. On the contrary, the significant drawback is that it uses a serial structure, which cannot train multiple trees simultaneously, making parallel processing difficult and increasing training time and parameter tuning difficulty (Liao et al., 2020; Nie et al., 2021). To gain a deeper understanding of predictive problems, including coagulant dosage prediction, by leveraging development and application of GBRT, readers are encouraged to refer to the following articles, such as Liao et al. (2020), Jiang et al. (2022), and Zhu et al. (2024).

2.6. Histogram gradient boosting (HistGBRT)

Histogram gradient boosting (HistGBRT) is a hybrid approach that combines gradient-based ensemble and binning-based methods to develop predictive modelling systems. Essentially, HistGBRT leverages the unique strengths of histogram-based plans to maximize the operational efficiency and information-adding potential of the system. This method enhances conventional boosting-based models by discretizing continuous feature variables into categorical *bins* (Hoang, 2025; Merabet et al., 2025; Nhat-Duc & Van-Duc, 2023). The partitioned parts represent the data distribution in a graph form, and each partition generates an important calculated metric. During the training process, each new decision tree in the overall model is built from previously trained trees. Within this category, newly added trees concentrate on minimizing predictive errors introduced by previous components (Fu, 2025; Merabet et al., 2025). This iterative mechanism gradually reduces the objective function, which represents the error magnitude of the entire ensemble system. Also, the objective function can be expressed as the following formula (8) (Hoang, 2025).

$$f_{Loss} = \frac{1}{N} \sum_{i=1}^N (y_i - t_i)^2 \quad (8)$$

where N = the number of data samples; y_i = the predicted values; and t_i = the actual values, respectively.

The significant strength of HistGBRT for predictive modelling is that it is much faster and uses less memory than traditional gradient boosting (e.g. XGBoost and LightGBM) on large datasets. The reason is that while a typical decision tree sorts and exhaustively examines all data to find the optimal split point, HistGBRT divides continuous variables into a predetermined number of *bins* and categorizes them (Guryanov, 2019; Marvin et al., 2023). In addition, the major drawback when applying HistGBRT is the loss of information due to discretization of data that occurs during the process of grouping data into *bins*. Because it forces continuous numerical data into specific *bins*, its accuracy may be reduced in predictive problems where very small numerical differences are important (Guryanov, 2019; Tamim Kashifi & Ahmad, 2022). More important papers and references on solving nonlinear problems, including coagulant dosage prediction, utilizing HistGBRT can be found in Guryanov (2019), Nhat-Duc and Van-Duc (2023), and Merabet et al. (2025).

2.7. Light gradient boosting machine (LightGBM)

Light gradient boosting machine (LightGBM) is a scalable computing platform that supports accelerated gradient-based ensemble learning. It integrates two optimization methods, including exclusive feature bundling (EFB) and gradient-based one-sided sampling (GOSS), into its structural framework for data processing and predictive modelling tasks (Fan et al., 2019; Guo et al., 2023). It implements a depth-limiting mechanism to address the potential overfitting problem that can arise from the generation of overly complex decision trees during the final node-centric growth process. This mechanism is a structural element that aims to optimize performance while minimizing the possibility of overfitting (Stefanovska & Pepelnjak, 2025). It focuses on reducing the expected value of a given objective function. The optimization criterion for this framework can be expressed in the following formula (9) (Y. Zhou et al., 2022).

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (9)$$

where $l(y_i, \hat{y}_i)$ = the loss function between the active and predicted criterion; $\sum_{i=1}^n l(y_i, \hat{y}_i)$ = the measure of how well the model fits the data; and $\Omega(f_k)$ = the regularization term for the k -th tree, respectively. The regularization (e.g. constraint) term can be set utilizing the formula (10) (Y. Zhou et al., 2022).

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda ||\omega||^2 \quad (10)$$

where, f_k = the k -th regression tree; T = the number of leaves in the tree; γ = the complexity penalty for each leaf; ω = the predictive values at leaves; $\|\omega\|^2$ = L2 norm of leaf weights; and λ = regularization parameter to shrink leaf scores, respectively.

The strength of LightGBM for solving predictive problems lies in the use of GOSS to focus on data with high gradients and sample only a small subset of data. Furthermore, it utilizes EFB to bundle variables, reducing computational complexity. As a result, it trains very quickly, and utilizes memory efficiently, even on large datasets. Explaining the weakness of LightGBM, in addition, because it adopts a leaf-wise growth method that immediately splits the node with the largest loss value, it is easy to experience overfitting, which is excessive optimization only for the training data, when the amount of data is small or the hyperparameter settings are inappropriate (Guo et al., 2023; K. Xu et al., 2023). There are many examples of LightGBM being commonly applied to non-linear predictive problems, including coagulant dosage prediction, in the previous literature, such as Fan et al. (2019), Guo et al. (2023), and Stefanovska and Pepelnjak (2025).

2.8. Natural gradient boosting (NGBoost)

Natural gradient boosting (NGBoost) is a boosting-based algorithm that goes beyond simple point estimation to calculate the entire probability distribution of predicted values (Duan et al., 2020). A key feature of NGBoost is that it generates distributional estimates rather than single predicted values. This allows for a more detailed analysis of possible scenarios, and more effectively reflects the probabilistic characteristics inherent in empirical data (Kavzoglu & Teke, 2022; S. Zhang et al., 2024). The adaptive framework of NGBoost for selecting optimization criteria provides compatibility with a variety of statistical distributions, improving efficiency across heterogeneous data environments. It utilizes the gradient-based ensemble approach to improve predictive accuracy by incrementally adding fundamental predictors over iterative training steps (Park et al., 2024; Rahman et al., 2025). Also, it can be updated based on stage t of the learning process through the following formula (11) (Rahman et al., 2025).

$$F_{t+1}(x) = F_t(x) + \eta h_t(x) \quad (11)$$

where, $F_t(x)$ = the current distribution parameter; $h_t(x)$ = weak learner trained on natural gradients of negative log-likelihood; and η = the learning rate, respectively. In addition, the coefficient adjustment protocol

within NGBoost is as the following formula (12) (Rahman et al., 2025).

$$\theta_{t+1} = \theta_t - \eta(F_t^{-1} \nabla L_t) \quad (12)$$

where, θ_t = the parameters of the predicted distribution at step t ; ∇L_t = the ordinary gradient of the loss; F_t^{-1} = the Fisher information matrix; $F_t^{-1} \nabla L_t$ = the gradient preconditioned by the Fisher geometry; and η = the learning rate, respectively.

The important strength of NGBoost is that it can estimate distribution parameters, such as the standard deviation and the mean, during the training process, so it provides a probability distribution of predicted values rather than a simple point estimate, allowing for the identification of confidence intervals. However, since NGBoost must perform complex matrix operations at each step to find the inverse matrix of the Fisher information matrix, and calculate the natural gradient during the parameter optimization process, it has the disadvantage of being significantly slower in learning speed and consuming more computing resources than high-performance boosting models, such as XGBoost and LightGBM (Chen et al., 2022; Kavzoglu & Teke, 2022). The diverse researches, including Duan et al. (2020), S. Zhang et al. (2024), and Rahman et al. (2025), provide excellent case studies on prediction and application issues (e.g. coagulant dosage) utilizing NGBoost.

2.9. eXtreme gradient boosting (XGBoost)

eXtreme gradient boosting (XGBoost) is an open-source library that optimizes the gradient boosting algorithm for parallel computing, maximizing performance and speed (Babajide Mustapha & Saeed, 2016; Sheridan et al., 2016). It is based on the gradient boosting decision tree framework, and has received considerable recognition for outstanding performance in various application areas. Also, it sequentially develops and improves weak learners, thereby creating strong learners (D. C. Feng et al., 2020; Osman et al., 2021). These fundamental classifiers set individual decision boundaries that aim to minimize a prior loss function, and are ultimately merged to produce the final decision boundary. Unlike other machine learning models, XGBoost includes a regularization factor within its objective function to prevent overfitting issues. The main objective function can be represented as the following formula (13) (Trizoglou et al., 2021).

$$O = \sum_{t=1}^n L(y_t, F(x_i)) + \sum_{k=1}^t R(f_k) + C \quad (13)$$

where, $\sum_{t=1}^n L(y_t, F(x_i))$ = the training loss; $\sum_{k=1}^t R(f_k)$ = regularization term at iteration k that controls the

complexity of trees; and C = the constant term, respectively. Also, the regularization component can be provided as the following formula (14) (Trizoglou et al., 2021).

$$R(f_k) = \alpha H + \frac{1}{2} \eta \sum_{j=1}^H W_j^2 \quad (14)$$

where, H = the number of leaves in the tree; W_j^2 = the weight of leaf j ; α = L1 regularization term (i.e. Lasso penalty); and η = L2 regularization term (i.e. ridge penalty). The total loss is chosen by the sum of individual loss components. Since the data in a decision tree is mapped to leaf nodes, the total loss can be calculated by aggregating the loss values at these leaf nodes (Kavzoglu & Teke, 2022; Trizoglou et al., 2021).

XGBoost has the advantage of being able to make more precise predictions than the conventional boosting-based models, and effectively suppressing overfitting because it includes a regularization (L1/L2) function that controls the complexity within itself, and utilizes the second derivative (Hessian) of the loss function to more accurately find the optimal split point. Judging from the weaknesses of XGBoost, however, there are numerous hyperparameters that must be tuned to achieve optimal performance, and the combinations between them are complex. Furthermore, model settings are challenging due to memory shortage issues when the data is extremely large, and the training process consumes significant system resources (i.e. memory/CPU) (Osman et al., 2021; Trizoglou et al., 2021). Readers can find various knowledge and information, including coagulant dosage prediction, on the application of XGBoost in the following papers and manuscripts: Sheridan et al. (2016), Osman et al. (2021), and Trizoglou et al. (2021).

2.10. Shapley Additive exPlanations (SHAP)

Shapley Additive exPlanations (SHAP) is an explainable AI (XAI) method that utilizes the Shapley value of game theory to explain the predictive results of complex machine learning and deep learning models. It quantifies the contribution of each feature to the predicted outcome of a specific dataset, and applies game theory to machine learning and deep learning, which fairly distributes the overall performance to each feature (Lundberg & Lee, 2017; Nohara et al., 2022). The development background of the SHAP method was proposed to address the problems of black box models, lack of consistency, and the need for a theoretical basis (Ekanayake et al., 2022; Lundberg & Lee, 2017). First, the problem with black box models is that high-performance models, such as deep learning and XGBoost, have complex internal structures,

making it difficult to understand why the specific results were obtained (Prendin et al., 2023). Next, the conventional method of feature importance has the inconsistent problem in which importance is distorted depending on changes in the model structure or dataset (Alizamir et al., 2024). Finally, it was developed to provide a theoretical basis, including reliable explanatory indicators based on mathematically proven fairness and consistency (Ekanayake et al., 2022; L. Feng et al., 2025; Lundberg & Lee, 2017).

One of the advantages of the SHAP method can be explained by its mathematical robustness, satisfying all theoretical properties of local correctness, omission-freeness, and consistency based on Shapley values. It is also useful not only for describing individual data (i.e. local), but also for analysing trends in the entire model (i.e. global). In addition, it can be applied to any machine learning and deep learning models, and also provides an optimized and fast algorithm (Alizamir et al., 2024; Prendin et al., 2023).

The weakness of the SHAP method can be categorized as follows. First, since the SHAP method must consider combinations of all possible variables, computation time increases exponentially as the number of data and features increases. Consequently, it costs high computational expenses (García & Aznarte, 2020). Next, if the correlation between the feature variables is ignored and there is a strong correlation between them, the contribution of feature variables may be distributed, leading to an interpretation that contains errors (Fouchal et al., 2025; Merabet et al., 2025). Finally, SHAP values represent contribution, not causality, so care should be taken not to over-interpret the numerical results as causal influence (García & Aznarte, 2020; Prendin et al., 2023). Readers can find more previous articles and references, including coagulant dosage prediction, to help them understand and utilize the SHAP method from Lundberg and Lee (2017), Ekanayake et al. (2022), and L. Feng et al. (2025).

2.11. Local interpretable model-agnostic explanations (LIME)

Local interpretable model-agnostic explanations (LIME) is based on the assumption that even if the overall model is complex, it can be approximated by a simpler model in a very narrow range. Therefore, rather than trying to explain the entire complex model, it is a method of creating a linear model that explains a local area by observing changes around the specific data to be interpreted (Kumarakulasingham et al., 2020; Zafar & Khan, 2021). It is developed to verify the model reliability by examining the basis of actual accuracy in the process of providing



Figure 1. Location map of *Taksebt* drinking water treatment plant.

a correct answer. In addition, for a neutral approach to model interpretation, a general tool is required that can explain any model as long as it has predictor and criterion values, even if the internal structure, such as model parameters, is not clearly known (Palatnik de Sousa et al., 2019; Peng et al., 2024). The significant advantage of the LIME method is that it is model-agnostic (i.e. flexible) and can be applied regardless of model or dataset types (C. Zhang & Liu, 2025). In addition, it is visualized in a form that is easy for humans to understand, allowing for intuitive interpretation (Gharehbaghi et al., 2026b). Finally, it is computationally efficient compared to the SHAP method, which considers combinations of all variables, and the results can be investigated easily (Tosan et al., 2025).

On the contrary, the weakness of the LIME method is that it is unstable because it only analyses the dataset surrounding the target of explanation, and even a small deviation from the range can drastically reduce the reliability of the explanation (Tosan et al., 2025). Also, since it trains by randomly sampling the surrounding dataset, it is dependent on data sampling, and therefore the explainable results may vary slightly each time whenever it is run (Y. Zhang et al., 2021). The previous articles and documents in the field of predictive issues (e.g. coagulant dosage prediction) depending on diverse machine learning, deep learning, and LIME method can be found from Kumarakulasinghe et al. (2020), C. Zhang and Liu (2025), and Gharehbaghi et al. (2026b).

3. Study area and data used

3.1. Operational process of the drinking water treatment plant

The present investigation was conducted at Taksebt drinking water treatment plant, Algeria (Figure 1), which has a production capacity of approximately 607,000 m³ per day (Fouchal et al., 2025). The selection of this particular case study water treatment plant is highly justified by the fact that, this region has a particular hydrological and water quality variability, rendering it a unique case study site for Mediterranean climate challenges. The available and stored raw water at Taksebt dam reservoir is originates from a reservoir heavily influenced by weather fluctuation. Under such circumstances, raw water quality is characterized by high peak turbidity and severe organic matter fluctuations. During the wet season (winter and spring), heavy rainfall events trigger intense soil erosion within the watershed, causing raw water turbidity to spike dramatically. Conversely, during the dry summer and autumn periods, elevated water temperatures and prolonged water residence times in the reservoir stimulate algal proliferation and organic degradation, leading to sharp, unpredictable fluctuations in dissolved organic matter. This highly dynamic and fluctuating nature of the raw water quality presents a major operational challenge for stable treatment, making the Taksebt plant an ideal and typical case study for validating advanced predictive coagulant models. It operates using a conventional coagulation-flocculation process, in which aluminium

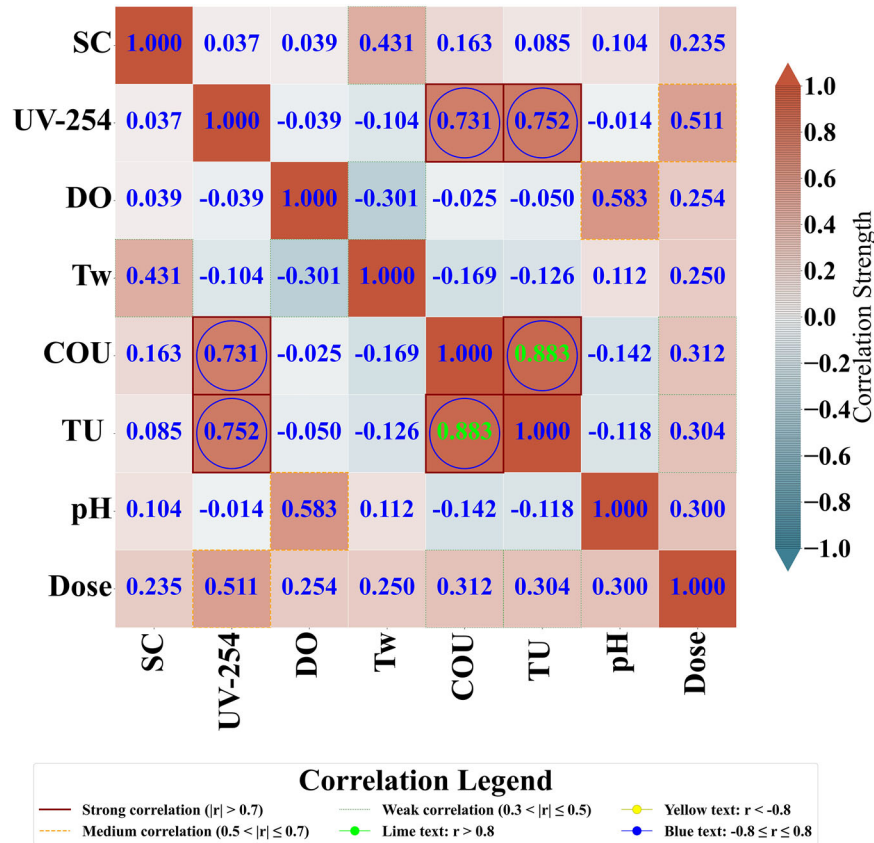


Figure 2. Pearson correlation heatmap.

sulphate ($\text{Al}_2(\text{SO}_4)_3 \cdot 18\text{H}_2\text{O}$) is applied as the primary coagulant, and has a storage capacity of 28,000 m^3 on site. To develop the predictive models, a comprehensive dataset was compiled, incorporating a range of raw water quality parameters, including: (i) water temperature (T_w), (ii) water pH, (iii) specific conductance (SC), (iv) water turbidity (TU), (v) water dissolved oxygen (DO), (vi) water Ultraviolet absorption (UV_{254}), and (vii) water Colour (COU), with a total of 1102 samples. The database was divided into two subsets: 70% (772 data points) for training and 30% (330 data points) for model validation. The statistical parameters for all variables are given in Table 1. Coefficients of correlation between coagulant dosage and raw water quality variables are depicted in Figure 2 in terms of heatmap.

3.2. Descriptive statistics of water quality variables

Table 1 shows the summary statistics for the water quality variables measured at the Taksebt Water Treatment Plant. The statistical values include mean (Mean), maximum (Max), minimum (Min), standard deviation (Std), coefficient of variation (C_V), Kurtosis, Skewness, correlation with coagulant dosage (R). The electrical conductivity (SC) values show low variability ($C_V \approx 0.086$), indicating a stable ionic composition in the raw water during

the monitoring period. Similarly, pH varies minimally ($C_V \approx 0.03$), indicating a well-buffered aquatic system. These stable values are likely to consistently contribute to estimating coagulant dosage. In contrast, UV_{254} , turbidity (TU), and colour (COU) exhibit substantially greater variability. UV_{254} has a coefficient of variation (C_V) of 0.787 in the training dataset, indicating substantial variation in dissolved organic matter concentration. Even greater variability is seen for COU ($C_V = 0.560$) and TU ($C_V = 1.659$), indicating significant fluctuations in particulate and colloidal matter in the raw water. Such high variation suggests nonlinear behaviour and complex interactions with coagulant requirements. Water temperature (T_w) exhibits moderate variability ($C_V \approx 0.21$ – 0.23), influenced by seasonal changes that can indirectly affect coagulation efficiency through changes in viscosity and reaction rates. Importantly, the target variable, coagulant dose (Dose), shows moderate variability ($C_V \approx 0.59$), indicating that prediction involves both stable and highly dynamic factors. The presence of high- C_V predictors (COU, TU, UV_{254}) alongside low- C_V variables (SC, pH) suggests a mixed linear and nonlinear system, justifying the use of both linear and more sophisticated ensemble learning models. Additionally, the similarity between training and validation statistics indicates that the dataset is statistically consistent and free

Table 1. Summary statistics of water quality variables at *Taksebt* station.

Variable	Subset	Mean	Median	Max	Min	Std	C _v	Kurtosis	Skweness	R
SC ($\mu\text{s}/\text{cm}$)	Train	487.935	482.750	639.500	424.500	41.739	0.086	−0.261	0.682	0.232
	Test	489.779	488.250	599.500	386.500	42.345	0.086	−0.474	0.505	0.243
	All Data	488.487	485.250	639.500	386.500	41.911	0.086	−0.336	0.627	0.235
UV ₂₅₄ (cm^{-1})	Train	5.114	4.600	35.150	0.041	4.025	0.787	9.949	2.429	0.520
	Test	4.710	4.550	20.300	0.041	3.341	0.709	4.070	1.381	0.490
	All Data	4.993	4.600	35.150	0.041	3.836	0.787	9.353	2.258	0.511
DO (mg/L)	Train	4.609	4.808	10.475	0.380	2.367	0.514	−1.137	−0.077	0.274
	Test	4.896	5.105	9.855	0.365	2.289	0.468	−1.088	−0.189	0.210
	All Data	4.695	4.890	10.475	0.365	2.347	0.514	−1.126	−0.113	0.254
Tw ($^{\circ}\text{C}$)	Train	17.344	16.575	29.800	8.250	3.693	0.213	0.752	0.972	0.217
	Test	17.190	16.250	29.850	11.200	3.897	0.227	1.109	1.203	0.323
	All Data	17.298	16.500	29.850	8.250	3.754	0.213	0.859	1.045	0.250
COU (HU)	Train	19.085	10.750	278.000	0.000	29.765	1.560	26.968	4.739	0.337
	Test	16.594	11.000	116.500	0.000	17.433	1.051	11.509	3.097	0.243
	All Data	18.339	11.000	278.000	0.000	26.693	1.560	30.609	4.903	0.312
TU (NTU)	Train	3.260	1.955	77.800	0.115	5.408	1.659	72.160	7.221	0.325
	Test	2.880	2.063	26.250	0.170	3.063	1.064	21.453	4.103	0.257
	All Data	3.146	1.990	77.800	0.115	4.829	1.659	81.408	7.452	0.304
pH (/)	Train	7.654	7.610	8.350	7.090	0.240	0.031	−0.284	0.535	0.292
	Test	7.669	7.635	8.290	7.130	0.229	0.030	−0.413	0.372	0.320
	All Data	7.658	7.615	8.350	7.090	0.237	0.031	−0.328	0.486	0.300
Dose (mg/L)	Train	11.028	10.000	30.000	3.000	6.467	0.586	0.565	0.997	1.000
	Test	10.836	10.000	30.000	3.000	6.430	0.593	1.134	1.187	1.000
	All Data	10.971	10.000	30.000	3.000	6.454	0.586	0.719	1.051	1.000

of major distributional shifts, supporting reliable model development. The linear correlation between raw water quality variables and coagulant dose was illustrated using a correlation matrix (Figure 2).

The correlation matrix indicates that the coagulant dose exhibits generally weak to moderate linear relationships with the raw water quality variables, suggesting that no variable significantly affects the coagulant dose variability. Overall, the correlation coefficient between coagulant dose and raw water quality variables ranged from $R \approx 0.235$ to ≈ 0.510 , and among all variables, UV₂₅₄ exhibited the strongest link with Dose ($R \approx 0.510$), indicating a moderate positive correlation and highlighting the importance of the absorption of ultraviolet light at a 254 nm wavelength in coagulant dose determination. The remaining variables, including COU (0.311), TU (0.304), pH (0.299), DO (0.254), T_w (0.249), and SC (0.2352), displayed weak positive correlations, reflecting limited individual linear influence on the applied dose. These relatively low coefficients suggest that coagulant dosing is governed by the combined, potentially nonlinear interaction among multiple parameters rather than by a single dominant factor.

Figure 3 illustrates the distributional behaviour of the raw water quality parameters and the corresponding coagulant dose, providing an integrated overview of seven input variables (SC, UV₂₅₄, DO, T_w , COU, TU, and pH) alongside the applied dosage. The combination of histograms, kernel density estimation (KDE), and superimposed normal distribution curves highlights clear differences in dispersion, symmetry, and shape across variables, reflecting the temporal heterogeneity of

raw water characteristics throughout the study period. To support a more rigorous statistical evaluation, supplementary analyses are presented in the Supplementary Material. These include a comparative statistical summary (Table S1) based on the Kolmogorov–Smirnov test, Jensen–Shannon distance, and Energy Distance metrics, together with three additional figures (Figures S1–S3) that depict cumulative distribution functions (CDF), probability density functions (PDF), and distance-based distribution comparisons, thereby offering a comprehensive assessment of distributional similarities and deviations among variables.

3.3. Input combination strategy for model development

Table 2 shows six hierarchical input configurations used consistently across all modelling methods, including MLR, RFR, AdaBoost, CatBoost, GBRT, HistGBRT, LightGBM, NGBoost, and XGBoost. The modelling framework follows a step-by-step feature elimination strategy:

- Scenario 1 (Full Model): All seven predictors (SC, UV₂₅₄, DO, T_w , COU, TU, pH)
- Scenario 2–6: Gradual removal of variables to test model sensitivity and reliability

This structured reduction has three main goals:

- Systematically removing predictors allows evaluation of each algorithm's stability and dependence on specific variables.

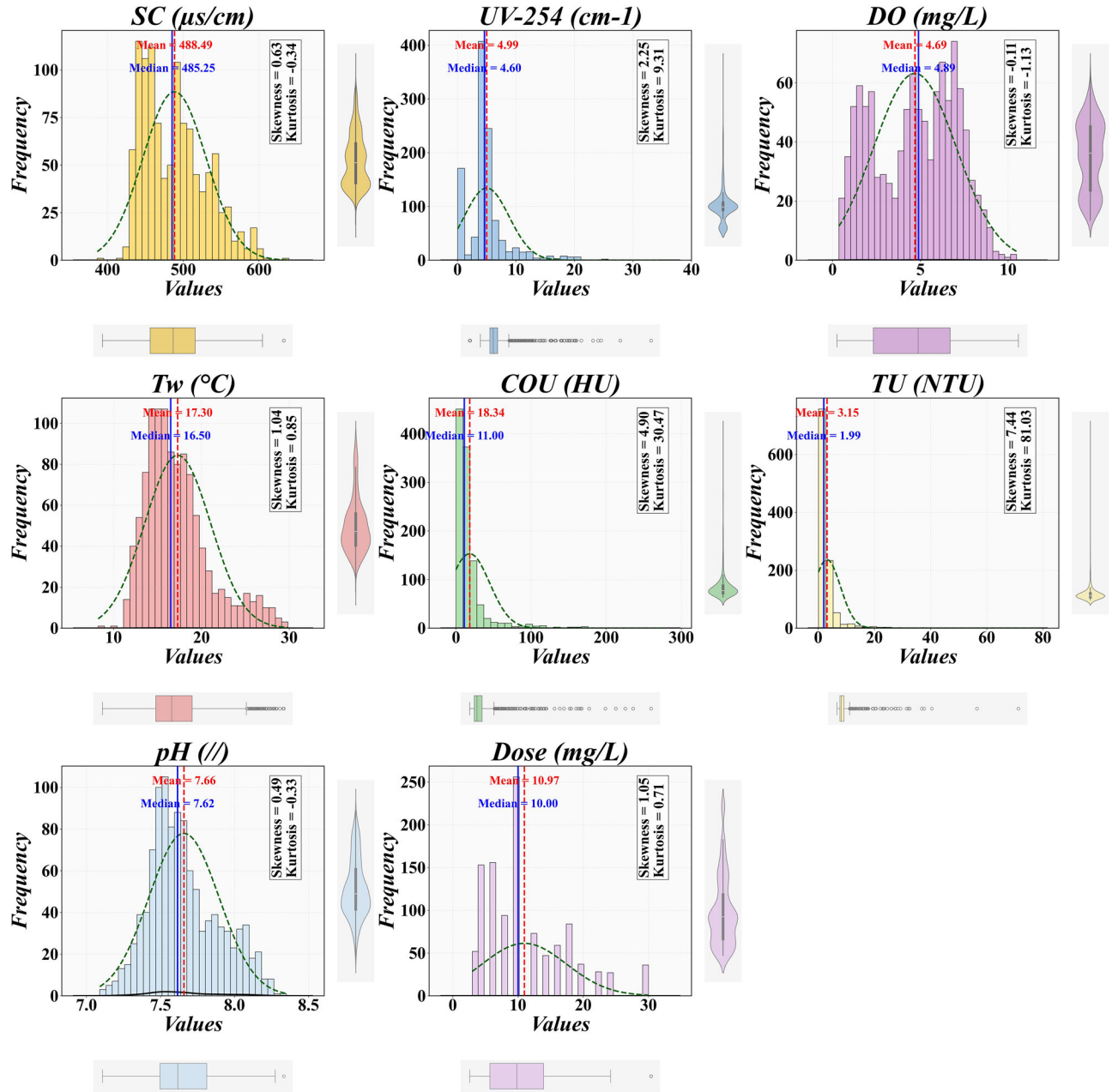


Figure 3. Distributional characteristics of the raw water quality variables and coagulant dose, with key descriptive statistical indicators.

- It helps identify the minimum set of variables needed for dependable dose prediction.
- Since all nine modelling frameworks use the same input scenarios, any performance differences can be attributed to the algorithms rather than data imbalance.

The stepwise reduction from seven to two variables also examines how boosting models handle limited information, compared with linear and bagging methods. Given the nonlinear and high variability characteristics noted in Table 1, particularly for TU and COU, effective boosting techniques are expected to exhibit greater

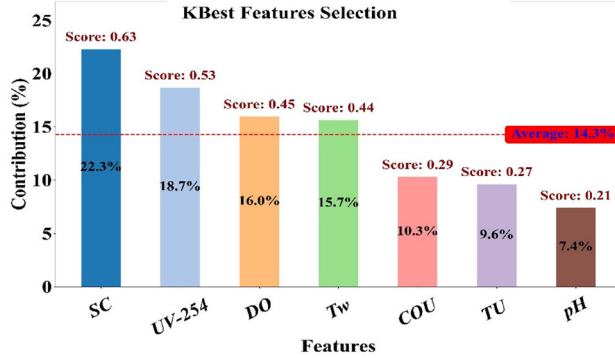
resilience when inputs are reduced. This experimental setup provides a fair way to assess whether CatBoost delivers the most reliable and explainable predictions across different levels of input complexity.

3.4. Mutual information-based feature importance analysis

Figure 4 displays the feature ranking using the KBest selection method based on mutual information scores. This method measures the nonlinear relationship between each predictor and the target variable (coagulant

Table 2. The input combinations of different models.

Models									Inputs	Output
MLR1	RFR1	AdaBoost1	CatBoost1	GBRT1	HistGBRT1	LightGBM1	NGBoost1	XGBoost1	SC, UV ₂₅₄ , DO, Tw, COU, TU, pH	Dose
MLR2	RFR2	AdaBoost2	CatBoost2	GBRT2	HistGBRT2	LightGBM2	NGBoost2	XGBoost2	SC, UV ₂₅₄ , DO, Tw, COU, TU	Dose
MLR3	RFR3	AdaBoost3	CatBoost3	GBRT3	HistGBRT3	LightGBM3	NGBoost3	XGBoost3	SC, UV ₂₅₄ , DO, Tw, COU	Dose
MLR4	RFR4	AdaBoost4	CatBoost4	GBRT4	HistGBRT4	LightGBM4	NGBoost4	XGBoost4	SC, UV ₂₅₄ , DO, Tw,	Dose
MLR5	RFR5	AdaBoost5	CatBoost5	GBRT5	HistGBRT5	LightGBM5	NGBoost5	XGBoost5	SC, UV ₂₅₄ , DO	Dose
MLR6	RFR6	AdaBoost6	CatBoost6	GBRT6	HistGBRT6	LightGBM6	NGBoost6	XGBoost6	SC, UV ₂₅₄	Dose

**Figure 4.** Histograms depicting features scores based on KBest features selection.

dose). Electrical conductivity (SC) is the most influential variable, accounting for 22.3% of the maximum mutual information score (0.63). This shows a strong link between ionic content and coagulant demand. UV₂₅₄ follows with an 18.7% contribution (score = 0.53), highlighting the significant impact of dissolved organic matter on coagulation processes. Dissolved oxygen (DO) and water temperature (T_w) contribute moderately (16.0% and 15.7%, respectively), indicating their indirect but important roles in reaction kinetics and organic matter behaviour. In contrast, colour (COU), turbidity (TU), and pH exhibit lower mutual information scores, with pH showing the least influence (7.4%, score = 0.21), suggesting it has a weaker relationship with dosage than the other factors. The average contribution threshold (14.3%) indicates that SC, UV₂₅₄, DO, and Tw are key predictors. This ranking supports the rationale for the progressive input-reduction strategy shown in Table 2 and facilitates future comparisons of model performance.

3.5. Accuracy evaluation criteria

The predictive performance of the developed models was assessed using four widely adopted statistical indicators: the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Nash-Sutcliffe Efficiency (NSE), and the Correlation Coefficient (R).

$$MAE = \frac{\sum_{i=1}^N |Dose_{pre,i} - Dose_{obs,i}|}{N} \quad (15)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Dose_{obs,i} - Dose_{pre,i})^2}{N}} \quad (16)$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^N (Dose_{obs,i} - Dose_{pre,i})^2}{\sum_{i=1}^N (Dose_{obs,i} - \overline{Dose_{obs}})^2} \right] \quad (17)$$

$$R = \left(\frac{\sum_{i=1}^N (Dose_{obs,i} - \overline{Dose_{obs}})(Dose_{pre,i} - \overline{Dose_{pre}})}{\sqrt{\sum_{i=1}^N (Dose_{obs,i} - \overline{Dose_{obs}})^2 \sum_{i=1}^N (Dose_{pre,i} - \overline{Dose_{pre}})^2}} \right) \quad (18)$$

In these expressions, $Dose_{obs,i}$ and $Dose_{pre,i}$ denote the observed and predicted coagulant dosages at the i^{th} time step, while N represents the total number of observations. The terms $\overline{Dose_{obs}}$ and $\overline{Dose_{pre}}$ correspond to the mean values of the observed and predicted coagulant dosages, respectively.

4. Results

This section evaluates the models developed to predict coagulant dosage at the Taksebt Water Treatment Plant. The analysis begins with a comparison of model performance, followed by a validation process, and an interpretability review. The main goal is to explore how predictive performance changes from linear regression to more complex ensemble and boosting methods as input complexity varies. At the same time, we assess robustness and interpretability to find the most dependable framework for optimizing operational dosage. A flowchart of all the modelling framework is depicted in Figure 5.

4.1. Comparative evaluation of machine learning models

The performance results of all models across different input configurations are shown in Table 3 and illustrated in Figures 2–6. We evaluated the models using correlation coefficient (R), Nash-Sutcliffe efficiency (NSE), root mean square error (RMSE), and mean absolute error (MAE), along with graphical and distribution-based analyses. We see clear improvements when moving from linear regression to ensemble and boosting methods. The

Table 3. Results of coagulant-dosage modelling at *Taksebt* water treatment plant.

Models	Training				Validation			
	<i>R</i>	NSE	RMSE (mg/L)	MAE (mg/L)	<i>R</i>	NSE	RMSE (mg/L)	MAE (mg/L)
MLR1	0.715	0.511	4.521	3.630	0.713	0.507	4.510	3.576
MLR2	0.715	0.511	4.522	3.634	0.712	0.506	4.514	3.584
MLR3	0.714	0.510	4.523	3.639	0.711	0.504	4.520	3.587
MLR4	0.709	0.503	4.556	3.657	0.709	0.500	4.541	3.610
MLR5	0.624	0.389	5.053	4.048	0.595	0.349	5.180	4.062
MLR6	0.556	0.310	5.370	4.142	0.555	0.299	5.373	4.070
RFR1	0.980	0.958	1.327	0.82	0.915	0.835	2.604	1.738
RFR2	0.980	0.956	1.349	0.839	0.915	0.835	2.612	1.754
RFR3	0.980	0.957	1.336	0.813	0.911	0.829	2.655	1.741
RFR4	0.977	0.953	1.399	0.851	0.906	0.820	2.725	1.796
RFR5	0.964	0.925	1.765	1.103	0.857	0.734	3.313	2.200
RFR6	0.925	0.850	2.500	1.710	0.762	0.574	4.191	3.012
AdaBoost1	0.863	0.706	3.506	3.069	0.849	0.667	3.703	3.229
AdaBoost2	0.844	0.678	3.667	3.179	0.832	0.640	3.853	3.337
AdaBoost3	0.854	0.695	3.571	3.106	0.836	0.646	3.820	3.328
AdaBoost4	0.834	0.675	3.686	3.101	0.817	0.632	3.896	3.287
AdaBoost5	0.793	0.588	4.146	3.601	0.758	0.519	4.452	3.830
AdaBoost6	0.753	0.539	4.390	3.657	0.742	0.521	4.444	3.683
CatBoost1	0.999	0.999	0.208	0.161	0.922	0.849	2.498	1.702
CatBoost2	0.999	0.999	0.229	0.180	0.918	0.842	2.550	1.752
CatBoost3	0.999	0.998	0.302	0.238	0.920	0.845	2.526	1.742
CatBoost4	0.998	0.996	0.397	0.303	0.913	0.834	2.613	1.754
CatBoost5	0.999	0.999	0.100	0.077	0.845	0.708	3.471	2.380
CatBoost6	0.971	0.939	1.593	1.078	0.756	0.559	4.263	3.060
GBRT1	0.999	0.999	0.138	0.100	0.909	0.826	2.682	1.804
GBRT2	0.999	0.999	0.170	0.124	0.907	0.821	2.713	1.881
GBRT3	0.999	0.999	0.194	0.145	0.903	0.815	2.764	1.881
GBRT4	0.999	0.998	0.282	0.204	0.893	0.796	2.897	2.006
GBRT5	0.998	0.996	0.428	0.316	0.860	0.739	3.282	2.306
GBRT6	0.970	0.939	1.595	1.015	0.750	0.544	4.337	3.049
HistGBRT1	0.984	0.967	1.177	0.742	0.904	0.817	2.745	1.897
HistGBRT2	0.984	0.967	1.179	0.760	0.903	0.816	2.754	1.949
HistGBRT3	0.982	0.964	1.229	0.806	0.902	0.813	2.774	1.943
HistGBRT4	0.979	0.957	1.343	0.897	0.896	0.803	2.852	1.974
HistGBRT5	0.959	0.917	1.866	1.256	0.852	0.726	3.362	2.365
HistGBRT6	0.895	0.798	2.906	2.098	0.760	0.573	4.197	3.037
LightGBM1	0.999	0.998	0.271	0.120	0.904	0.816	2.755	1.906
LightGBM2	0.999	0.997	0.336	0.158	0.896	0.803	2.852	1.992
LightGBM3	0.999	0.997	0.346	0.172	0.896	0.801	2.867	1.964
LightGBM4	0.998	0.996	0.423	0.242	0.889	0.789	2.951	2.029
LightGBM5	0.993	0.986	0.764	0.486	0.838	0.696	3.538	2.466
LightGBM6	0.954	0.907	1.967	1.371	0.746	0.541	4.351	3.140
NGBoost1	0.931	0.858	2.437	1.832	0.898	0.799	2.878	2.263
NGBoost2	0.927	0.851	2.497	1.906	0.896	0.794	2.915	2.304
NGBoost3	0.924	0.845	2.544	1.938	0.896	0.793	2.922	2.302
NGBoost4	0.924	0.846	2.538	1.947	0.895	0.793	2.920	2.323
NGBoost5	0.898	0.801	2.881	2.150	0.866	0.745	3.241	2.540
NGBoost6	0.846	0.710	3.481	2.559	0.805	0.646	3.822	2.862
XGBoost1	0.997	0.994	0.484	0.347	0.916	0.839	2.577	1.744
XGBoost2	0.997	0.994	0.505	0.363	0.912	0.831	2.640	1.804
XGBoost3	0.996	0.992	0.563	0.401	0.910	0.827	2.670	1.831
XGBoost4	0.995	0.989	0.663	0.477	0.900	0.810	2.799	1.898
XGBoost5	0.990	0.979	0.936	0.659	0.869	0.754	3.183	2.218
XGBoost6	0.955	0.907	1.967	1.339	0.776	0.591	4.103	2.940

sensitive to feature reduction, particularly with minimal input configurations. Boosting algorithms achieve the most notable improvements in predictive performance. Gradient-boosting variants typically achieve validation *R* values above 0.90 and NSE values above 0.80 in the most comprehensive scenarios. Error rates drop significantly compared to both MLR and AdaBoost, reflecting

a better ability to model nonlinear relationships among water quality variables.

Figures 6 and 7 summarize this trend in improvement. Error metrics (RMSE and MAE) significantly decrease when moving from linear regression to advanced boosting methods, while *R* and NSE rise accordingly. Although several gradient-boosting algorithms

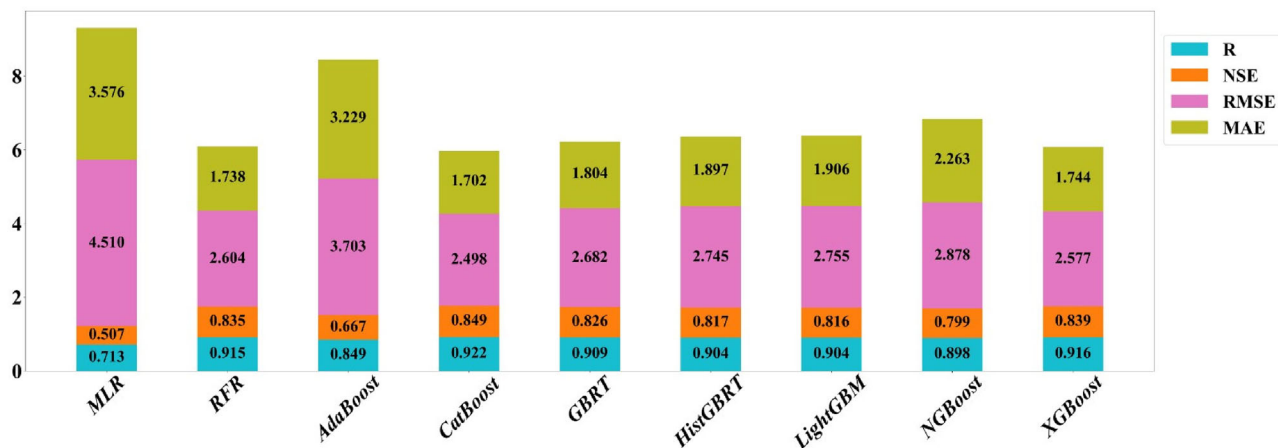


Figure 6. Bar plots of the metrics computed for the validation stage.

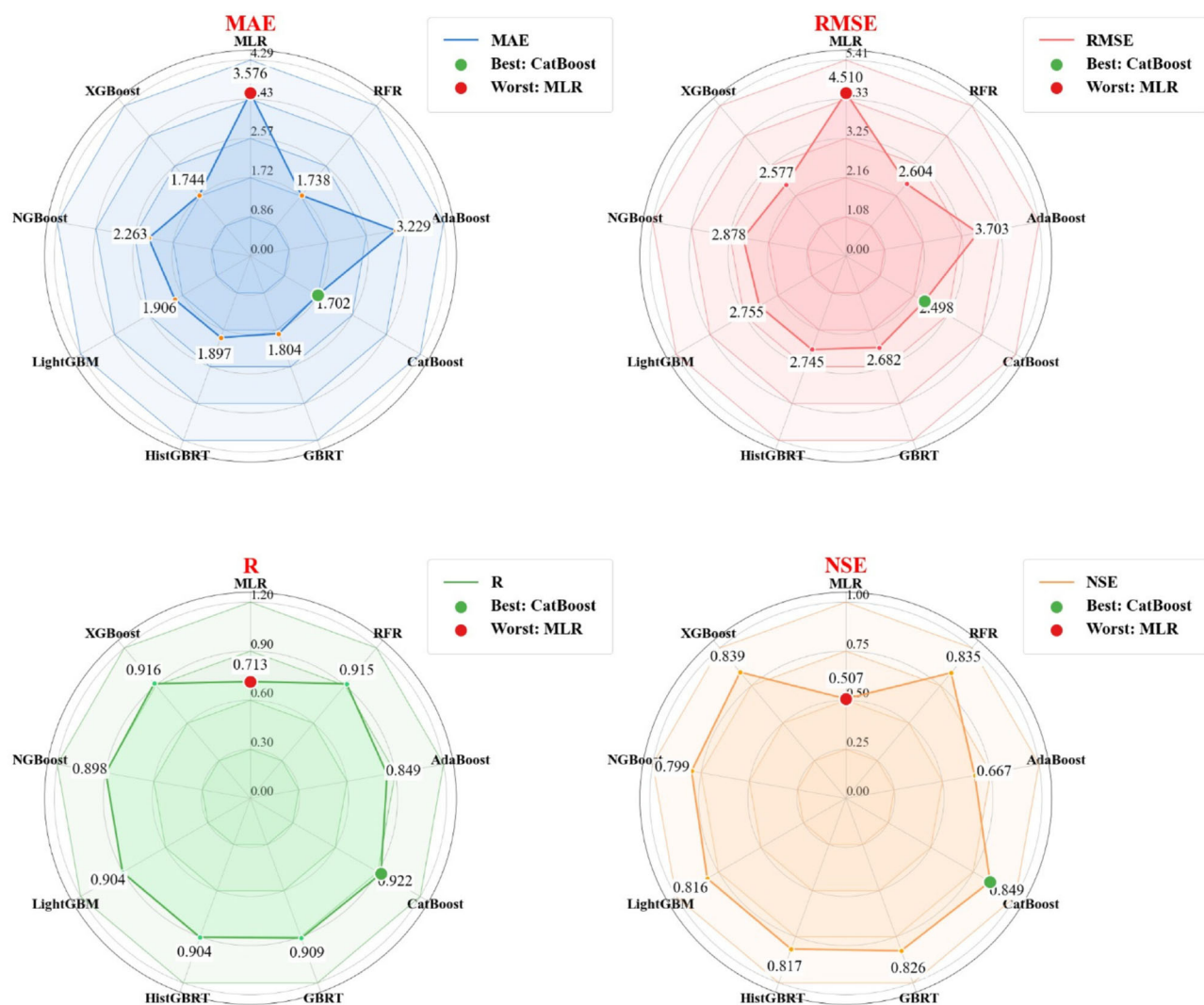


Figure 7. Radar plots illustrating performance for the validation stage.

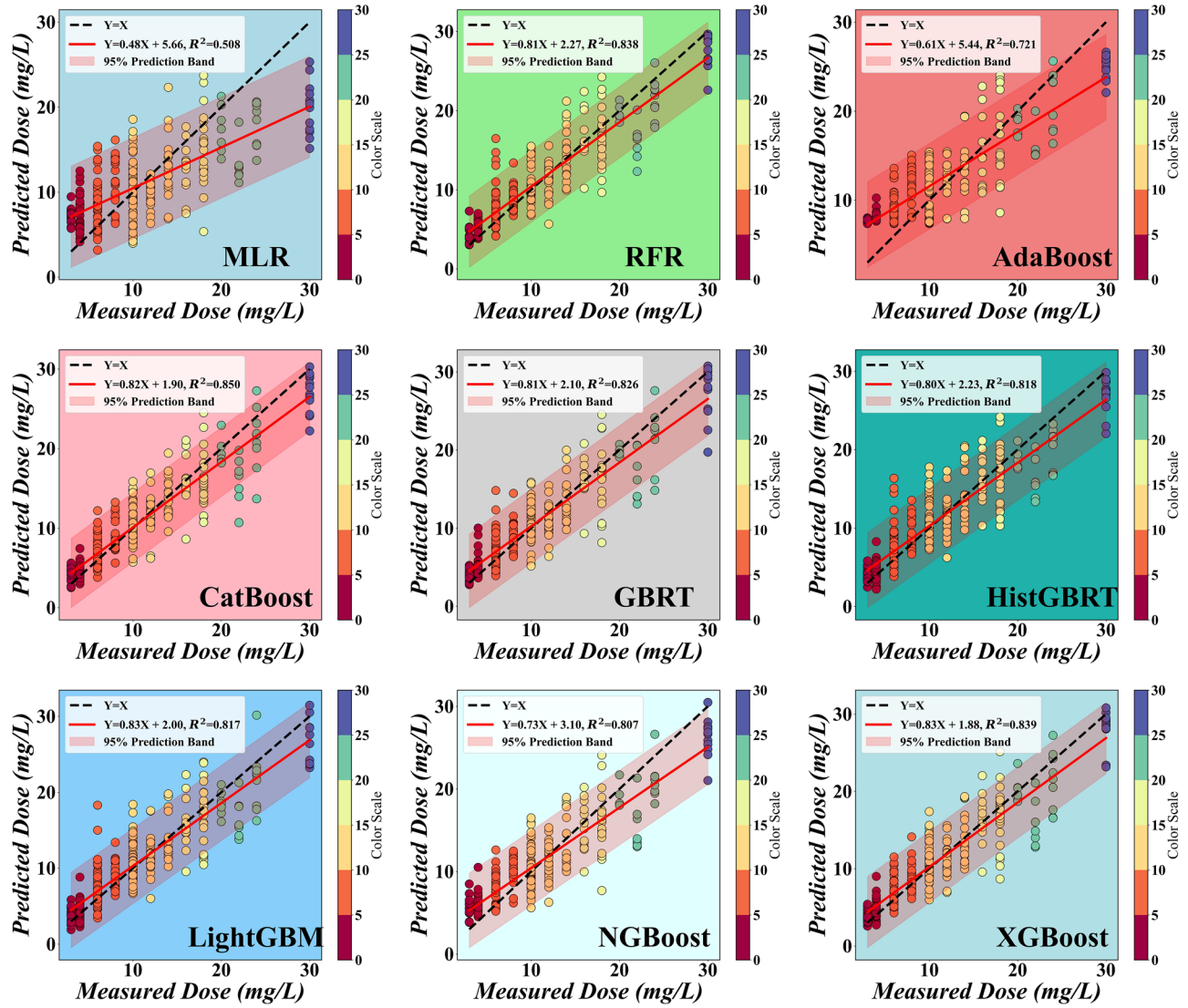


Figure 8. Scatterplots of measured versus predicted coagulant dosage using the best models for the validation stage.

perform well, CatBoost stands out for its balanced performance. It achieves the highest validation correlation ($R = 0.922$) and NSE (0.849), along with relatively low RMSE (2.498 mg/L) and MAE (1.702 mg/L). The scatter plots in Figure 8 support these findings. Linear regression shows noticeable spread around the 1:1 reference line, especially at higher dosage values, indicating consistent deviation under extreme conditions. Ensemble and boosting methods progressively tighten prediction spread. CatBoost is closest to the 1:1 line and has the narrowest 95% prediction band, showing better generalization and less bias across the operational range. The distribution comparisons in Figure 9 provide further insight into model reliability. While linear regression compresses the variability of predicted doses and underestimates extremes, boosting algorithms

better maintain the statistical structure of the measured distribution. The boxplot and violin plots show that CatBoost accurately reflects the central tendency, interquartile range, and overall spread of observed dosage values, indicating better representation of the underlying process variability. Finally, the Taylor diagram in Figure 10 provides a statistical comparison of the standard deviation, correlation, and centred root-mean-square deviation. CatBoost's proximity to the reference point indicates an optimal balance between high correlation and appropriate variance representation. Other gradient-boosting methods cluster near the reference area, whereas MLR is farther away, indicating its weaker performance. Overall, the combined evidence from numerical metrics, graphical patterns, distribution similarity, and the Taylor diagram consistently shows that boosting algorithms outperform linear regression.

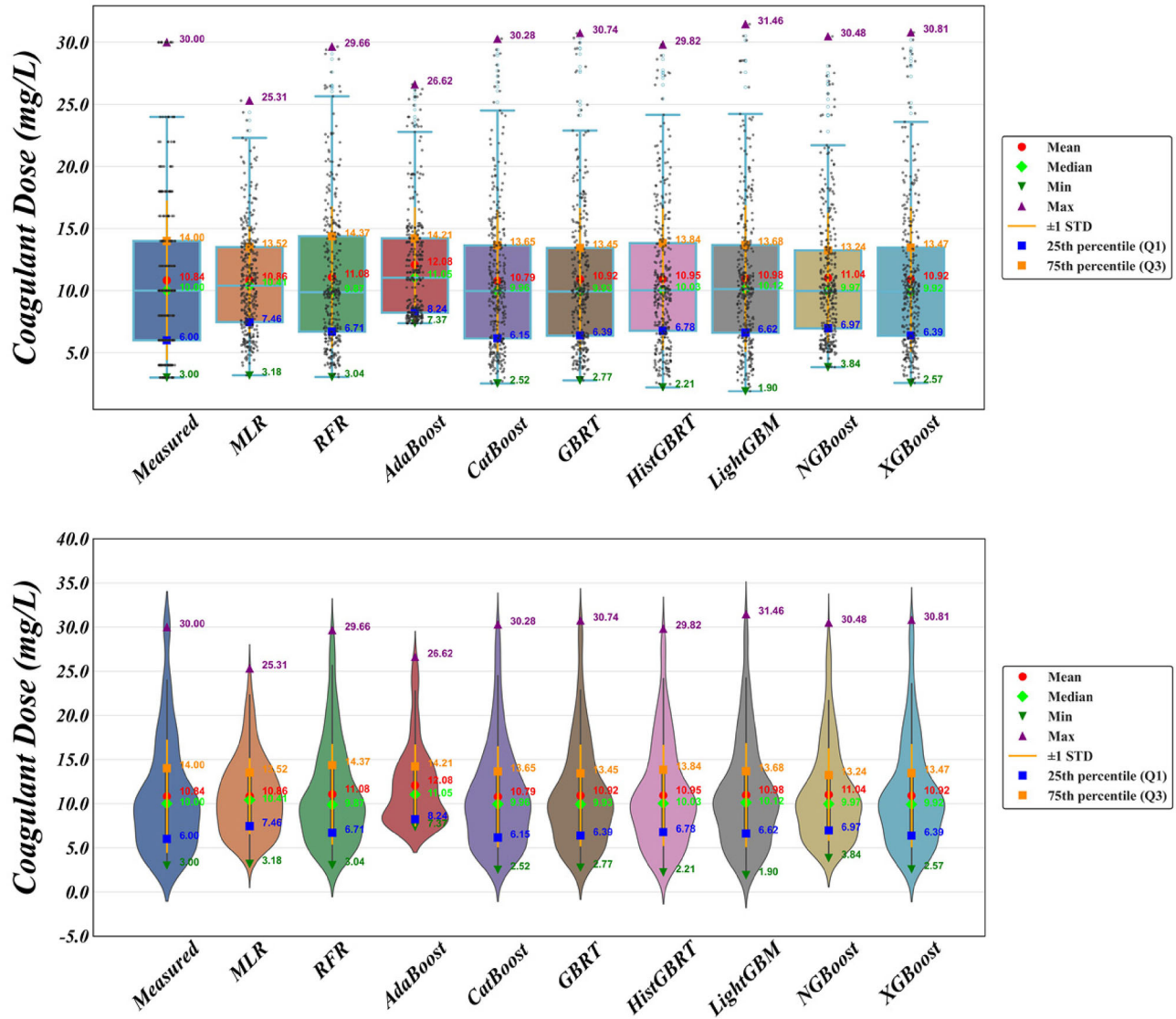


Figure 9. Boxplot and violinplot for coagulant dosage prediction using the best models.

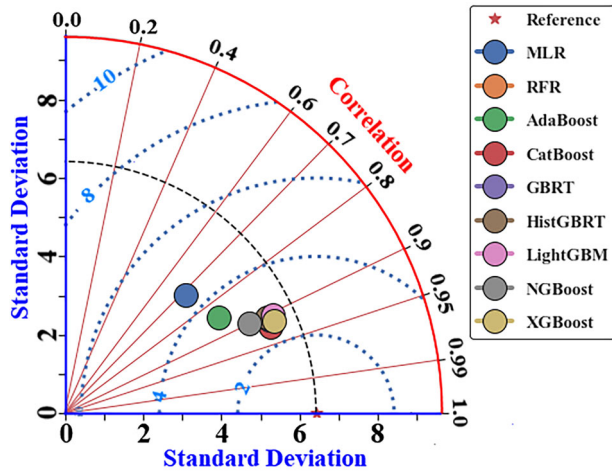


Figure 10. Taylor diagram for coagulant dosage prediction using the best models.

Among them, CatBoost exhibits the most reliable performance across evaluation metrics, providing a strong foundation for clear and robust predictions of coagulant dosage.

To further evaluate the model's robustness, predictive performance across the six hierarchical input scenarios was compared. As the number of input variables decreased from Scenario 1 (seven variables) to Scenario 6 (two variables), all models experienced some deterioration in predictive accuracy. However, the magnitude of degradation differed considerably among algorithms. For example, the validation NSE of MLR decreased from 0.507 to 0.299 (approximately 41%), while RMSE increased from 4.510 to 5.373 mg/L (about 19%), indicating a strong dependence on the complete predictor set. A similar trend was observed for AdaBoost, whose NSE decreased by approximately 22.5% (0.667 to 0.521),

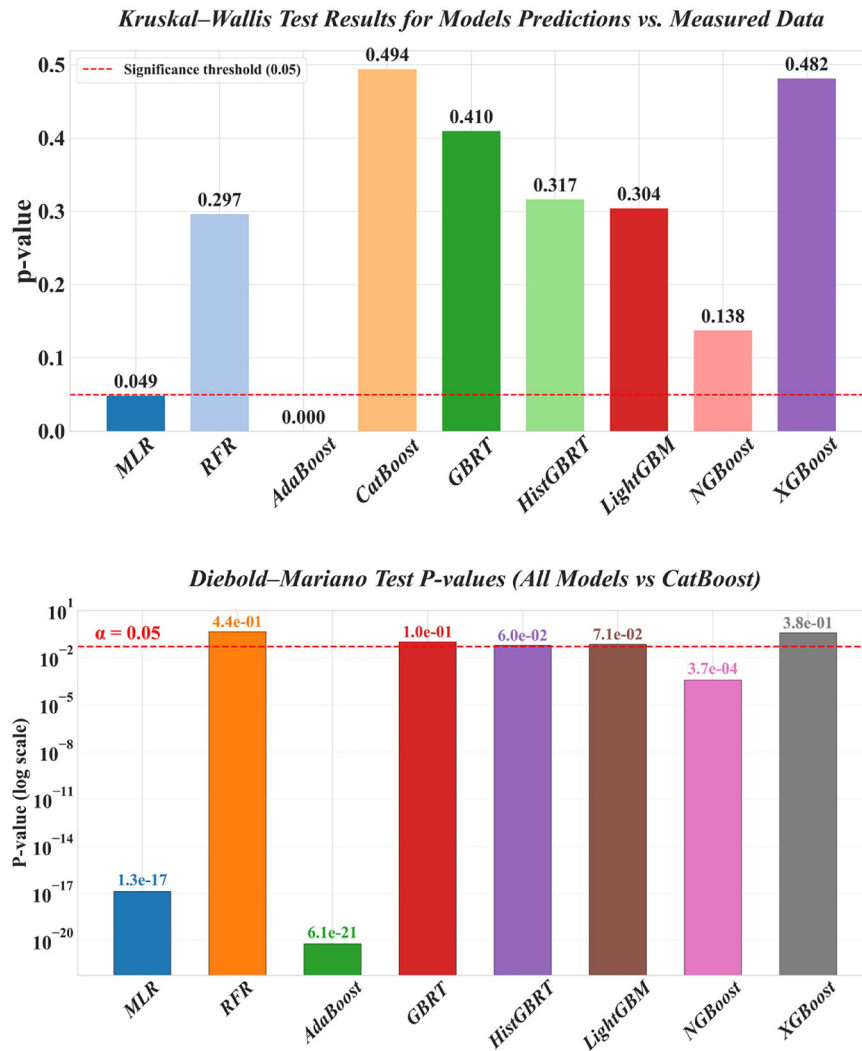


Figure 11. Statistical comparison of model predictions using Kruskal–Wallis and Diebold–Mariano tests.

accompanied by an increase in RMSE from 3.703 to 4.444 mg/L. In contrast, CatBoost maintained comparatively stable performance under feature reduction, with validation NSE decreasing from 0.849 to 0.559 (34%) and RMSE increasing from 2.498 to 4.263 mg/L. Although all boosting algorithms exhibited some loss of accuracy, CatBoost consistently provided the highest prediction accuracy across all six scenarios. Likewise, XGBoost, GBRT, and LightGBM demonstrated greater resilience to feature reduction than MLR and AdaBoost, confirming the superior ability of boosting algorithms to preserve predictive performance even when only limited water-quality information is available. These findings indicate that boosting models are considerably more tolerant of incomplete feature information than conventional linear regression models, making them more suitable for practical water treatment applications where some measurements may be unavailable or delayed.

4.2. Statistical significance and forecast comparison tests

To further verify the reliability of model predictions beyond standard performance metrics, we conducted non-parametric and forecast comparison tests. The results of the Kruskal–Wallis test and Diebold–Mariano (DM) test are shown in Figure 11 and Tables 4 and 5. The Kruskal–Wallis test checks for statistically significant differences between measured and predicted dosage distributions at the 95% confidence level. As shown in Table 4, MLR ($p = 0.0488$) and AdaBoost ($p \approx 0.0000$) show significant differences from measured data ($p < 0.05$), leading to the rejection of the null hypothesis. This confirms that these models do not accurately reproduce the observed dosage distribution. In contrast, CatBoost ($p = 0.4945$), GBRT ($p = 0.4099$), XGBoost ($p = 0.4817$), and other gradient boosting models show p -values well above the significance threshold. Therefore,

Table 4. *P*-values of Kruskal-Wallis test at 95% significance level.

Models	Coagulant Dosage (mg/L)		
	Test statistic (H)	<i>p</i> -value	H_0
MLR	3.8808	0.0488	Reject $H_0 \rightarrow$ Significant
RFR	1.0894	0.2966	Fail to reject (Not significant)
AdaBoost	19.7666	<0.0000	Reject $H_0 \rightarrow$ Significant
CatBoost	0.4666	0.4945	Fail to reject (Not significant)
GBRT	0.6792	0.4099	Fail to reject (Not significant)
HistGBRT	1.0031	0.3166	Fail to reject (Not significant)
LightGBM	1.0545	0.3045	Fail to reject (Not significant)
NGBoost	2.202	0.1378	Fail to reject (Not significant)
XGBoost	0.495	0.4817	Fail to reject (Not significant)

we cannot reject the null hypothesis, indicating no significant difference between predicted and observed distributions. Among all models, CatBoost has one of the highest *p*-values, suggesting it aligns closely with measured data.

Figure 11 confirms these findings, with boosting models safely above the significance threshold, whereas MLR and AdaBoost fall below it. To compare predictive accuracy between CatBoost and other algorithms, we applied the Diebold-Mariano test (Table 5). The DM test assesses whether two competing models have significantly different forecast errors. Results show that CatBoost significantly outperforms MLR ($p = 1.32\text{e-}17$), AdaBoost ($p = 6.08\text{e-}21$), and NGBoost ($p = 3.7\text{e-}04$), as their *p*-values are well below 0.05 and the test statistics indicate CatBoost's superiority. However, when compared with RFR, GBRT, HistGBRT, LightGBM, and XGBoost, we find no significant differences ($p \geq 0.05$). This indicates that, although several gradient-boosting methods are competitive, CatBoost often matches or exceeds their forecasting accuracy. Together, the Kruskal-Wallis and Diebold-Mariano test results support the strength of CatBoost. It not only achieves high predictive accuracy and low error metrics, but also shows consistency with observed data and statistically validated superiority over weaker models.

4.3. SHAP interpretability analysis

Global feature importance results (Figure 12) indicate that UV_{254} exhibits the highest mean SHAP value across nearly all evaluated models. In the best performing model, CatBoost, UV_{254} clearly emerges as the most influential predictor, accounting for 23.6% of the total contribution. It is followed by SC (18.8%), T_w (18.5%), DO (16%), and pH (11%), while TU (6.8%) and COU (5.3%) exhibit comparatively smaller impacts. A similar ranking pattern is observed across most of the other models, including MLR, RFR, AdaBoost, GBRT, NGBoost, and XGBoost, UV_{254} also remains the most influential variable, consistently presenting the highest mean SHAP contribution. However, a different pattern

is observed for HistGBRT and LightGBM, where SC slightly surpasses UV_{254} and becomes the leading contributor. Despite this variation, UV_{254} remains among the top-ranked predictors in all models, confirming its overall robustness and dominant influence in coagulant dosage prediction. The SHAP summary plots reveal a broader distribution of SHAP values for CatBoost compared to MLR, highlighting their superior capability to capture nonlinear relationships and threshold-dependent behaviours. In most architectures, higher UV_{254} values are associated with positive SHAP contributions, indicating an increase in the predicted coagulant dosage. Conversely, lower UV_{254} values correspond to negative SHAP contributions, reflecting a reduced predicted dosage.

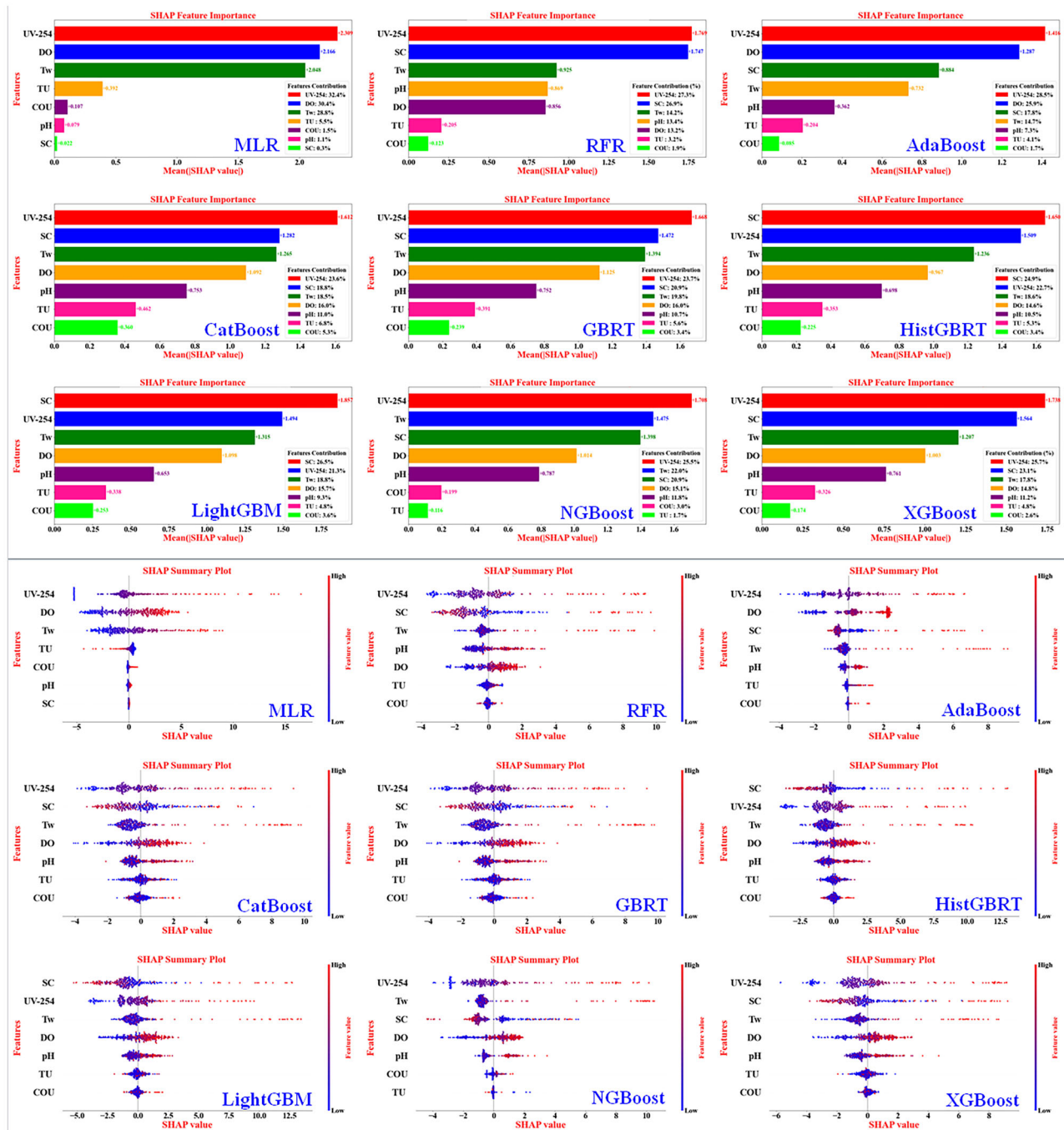
The dominance of UV_{254} is also physically meaningful from a coagulation perspective. UV_{254} is widely recognized as a surrogate indicator of dissolved aromatic organic matter, particularly humic and fulvic substances, which are major contributors to natural organic matter (NOM) in surface waters. These compounds possess negatively charged functional groups that consume coagulant during charge neutralization and sweep flocculation processes. Consequently, increases in UV_{254} generally require higher aluminium sulphate dosages to destabilize dissolved organic matter and produce effective floc formation. Therefore, the SHAP results are not only statistically consistent but also agree with the established physicochemical mechanisms governing coagulation processes. This agreement increases confidence that the CatBoost model captures meaningful process relationships rather than purely data-driven correlations.

4.4. Local explanation analysis using SHAP and LIME

To complement the global SHAP analysis, we applied local explanation techniques to select representative samples. These visualizations show how individual feature values contribute to specific dosage predictions, providing clear insights at the sample level. Figure 13(a–f) shows single-instance SHAP contribution plots for six representative samples covering low, medium, and high dosage conditions. The results reveal that prediction behaviour is influenced by dynamic interactions rather than fixed linear relationships. For lower-predicted-dosage cases (e.g. Samples 20 and 280), lower UV_{254} and moderate SC values result in lower predicted values, whereas temperature and pH exert secondary adjustments. Conversely, in higher-dose samples (e.g. Sample 100 and Sample 250), elevated UV_{254} , higher SC, and higher temperature values increase the predictions. Importantly, SHAP contributions exhibit physically consistent behaviour:

Table 5. *P*-values of Diebold-Mariano test at 95% significance level.

Comparison models	Test statistic <i>S</i> ₀	<i>p</i> -value (<i>P</i> ₀)	Result	Decision
MLR	−9.0457	1.320e-17	$S_0 < 0, P_0 < 0.05$, CatBoost is better.	Reject
RFR	−0.7713	0.4411	$P_0 \geq 0.05$, no significant difference.	Fail
AdaBoost	−10.0624	6.084e-21	$S_0 < 0, P_0 < 0.05$, CatBoost is better.	Reject
GBRT	−1.6429	0.1014	$P_0 \geq 0.05$, no significant difference.	Fail
HistGBRT	−1.8858	0.0602	$P_0 \geq 0.05$, no significant difference.	Fail
LightGBM	−1.8133	0.0707	$P_0 \geq 0.05$, no significant difference.	Fail
NGBoost	−3.5962	0.0004	$S_0 < 0, P_0 < 0.05$, CatBoost is better.	Reject
XGBoost	−0.8868	0.3758	$P_0 \geq 0.05$, no significant difference.	Fail

**Figure 12.** SHAP global feature importance results for coagulant dosage prediction: bar plots for global feature importance upper panel and SHAP summary plots lower panel.

higher UV_{254} values generally increase the predicted coagulant demand. Elevated SC levels also positively contribute, reflecting ionic strength effects. Temperature shows nonlinear interaction patterns based on concentration. Minor predictors (e.g. TU and COU) have localized but not dominant effects.

Figure 14(a–f) presents LIME-based local explanations for the same samples. LIME explanations confirm the directional influence identified by SHAP, further enhancing model interpretability. While LIME provides rule-based threshold interpretations (e.g. $UV_{254} > \text{threshold}$ leads to increased dosage), SHAP quantifies additive contributions. Notably, CatBoost's explanations appear smoother and more stable across samples compared to other models, avoiding abrupt behaviour based on thresholds. The alignment between SHAP and LIME interpretations indicates that CatBoost maintains consistent local-explanation patterns without relying excessively on misleading splits or extreme partitions. The combination of stable global importance (Figure 12), consistent local explanations (Figures 13 and 14), distributional agreement (Figure 9), and statistical validation (Figure 11) indicates that CatBoost offers both predictive accuracy and interpretable decision logic. This integrated interpretability framework supports the use of CatBoost for real-world water treatment plant operations, where transparent decision support is crucial.

5. Discussion

The analysis of results clearly revealed that boosting models, particularly CatBoost, represent a clear advancement in coagulant dose prediction. This study shares similarities with prior studies in using parameters like turbidity, pH, and temperature. However, differences exist. Many earlier papers (pre-2020) used ANFIS or basic artificial neural networks (ANNs).

In other words, unlike this study that focused on the application of explainable artificial intelligence (XAI) and ensemble boosting techniques, earlier studies often relied on simpler approaches, such as ANFIS (Wu & Lo, 2008), ANN (Tochio et al., 2023), SVM (Tahraoui et al., 2021) or more advanced integrated ones like hybrid ELM-Bat algorithm (Boumezbeur et al., 2023) for the prediction of coagulant dose in water treatment plants, without post-hoc interpretability. To put it more simply, their approach lacked the interpretability features essential for operational deployment. On the contrary, recent works published in 2025 and 2026 shifted their modelling approach towards the employment of ensemble models, such as boosting methods, and towards gaining the advantages of different XAI techniques, such as SHAP, similar to

this study. The majority of these studies reported better performance of ensemble models over individual ML models, but with different inputs and contexts (Fouchal et al., 2025; W. Zhao et al., 2026).

To fully confirm the international validity and technical advantages of the proposed framework, thorough horizontal benchmarking against various global water matrices and regional research is necessary. Internationally, coagulant dose prediction accuracy highly depends on the stability of the water source, modelling technique, and regional climatic factors. For example, in East Asian plants dealing with real-world scale data sets, alternative parametric methods such as Lasso-regularized polynomial multiple linear regression (PMLR) have yielded an R^2 of 0.951 from long-term operation data, showing that regularization of linear variables helps to avoid overfitting provided source waters stay within localized distributional limits (J. An et al., 2025). At the same time, recent paradigm changes in modelling employ NSGA-II dual-objective optimization methods in adaptive stacking models with multi-layer perceptron and Gaussian process predictors, yielding predictions with accuracy above an R^2 of 0.990 on particular surface water data sets (W. Zhao et al., 2026). In addition, predictions made for the water management systems of regions involving the processing of volatile surface water that contains high amounts of organic content frequently have poor Nash-Sutcliffe efficiencies ($NSE < 0.78$) owing to the quick and drastic seasonal changes in the amount of organic material (Fouchal et al., 2025). On the other hand, our CatBoost model, when applied to the volatile surface waters of the Taksebt station in Algeria, attains an impressive validation correlation coefficient of ($R = 0.922$) and a relatively high efficiency ($NSE = 0.849$) in configuration #1. This shows that sequential boosting models have excellent cross-region adaptability and generalizability than local-based models.

For example, L. Feng et al. (2025) used automated machine learning with RFR and SHAP for interpretable coagulant dosage prediction. Although the boosting techniques performed well in predicting coagulant dosage, the ensemble bagging RFR model surpassed the gradient boosting trees model, achieving a high R^2 (0.96) and a low RMSE (0.89). Despite the fact that different input parameters were used, their SHAP analysis identified SC, ammonia nitrogen, chemical oxygen demand, and Tw of raw water as key factors affecting coagulant dosage, which is somewhat parallel to the results of the SHAP analysis in this study. In another study, some researchers examined several deep learning models along with SHAP as an XAI technique for predicting coagulant dosage in an Algerian plant (Fouchal et al., 2025).

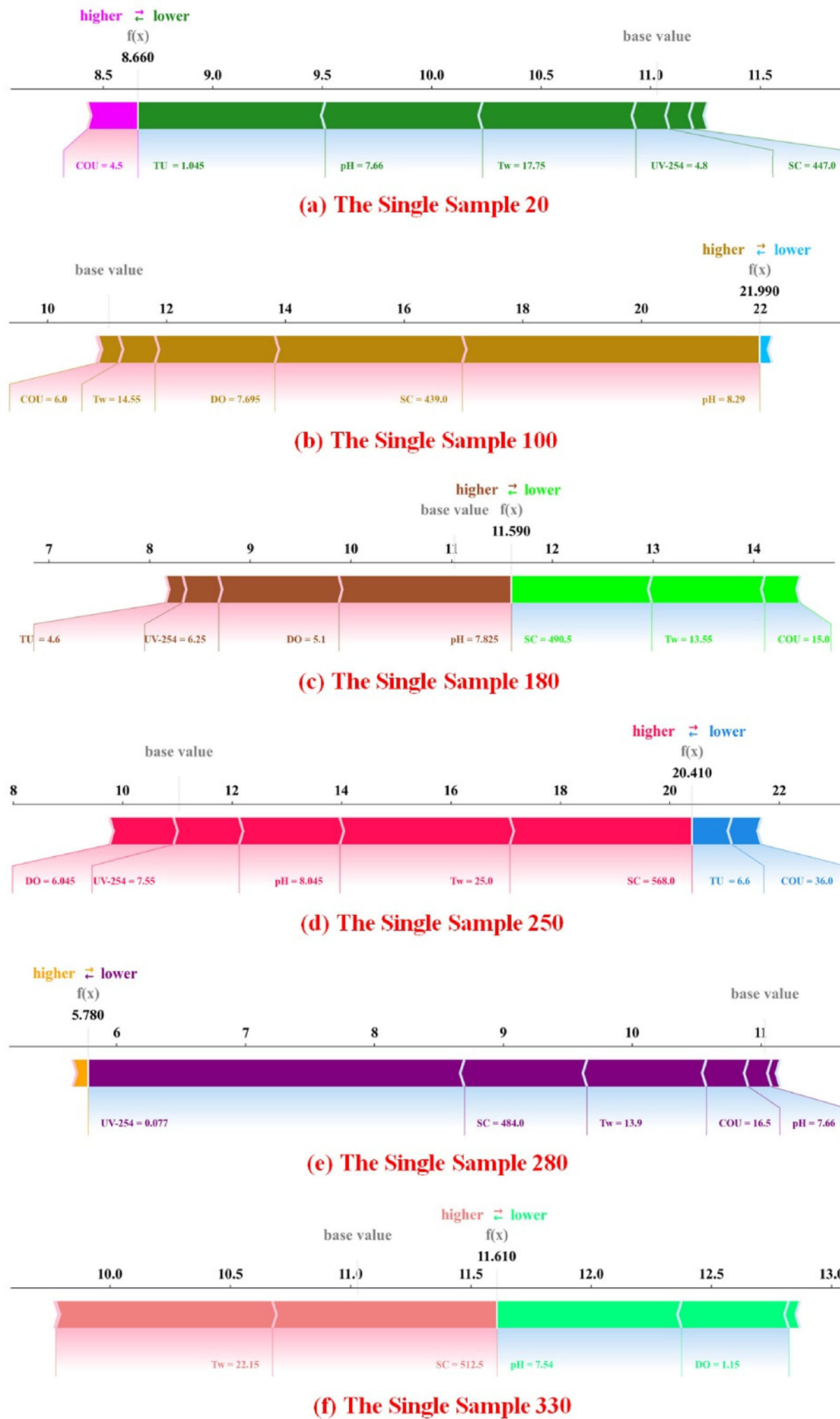


Figure 13. SHAP force plots for local interpretability of individual coagulant dosage predictions across six representative samples (a–f).

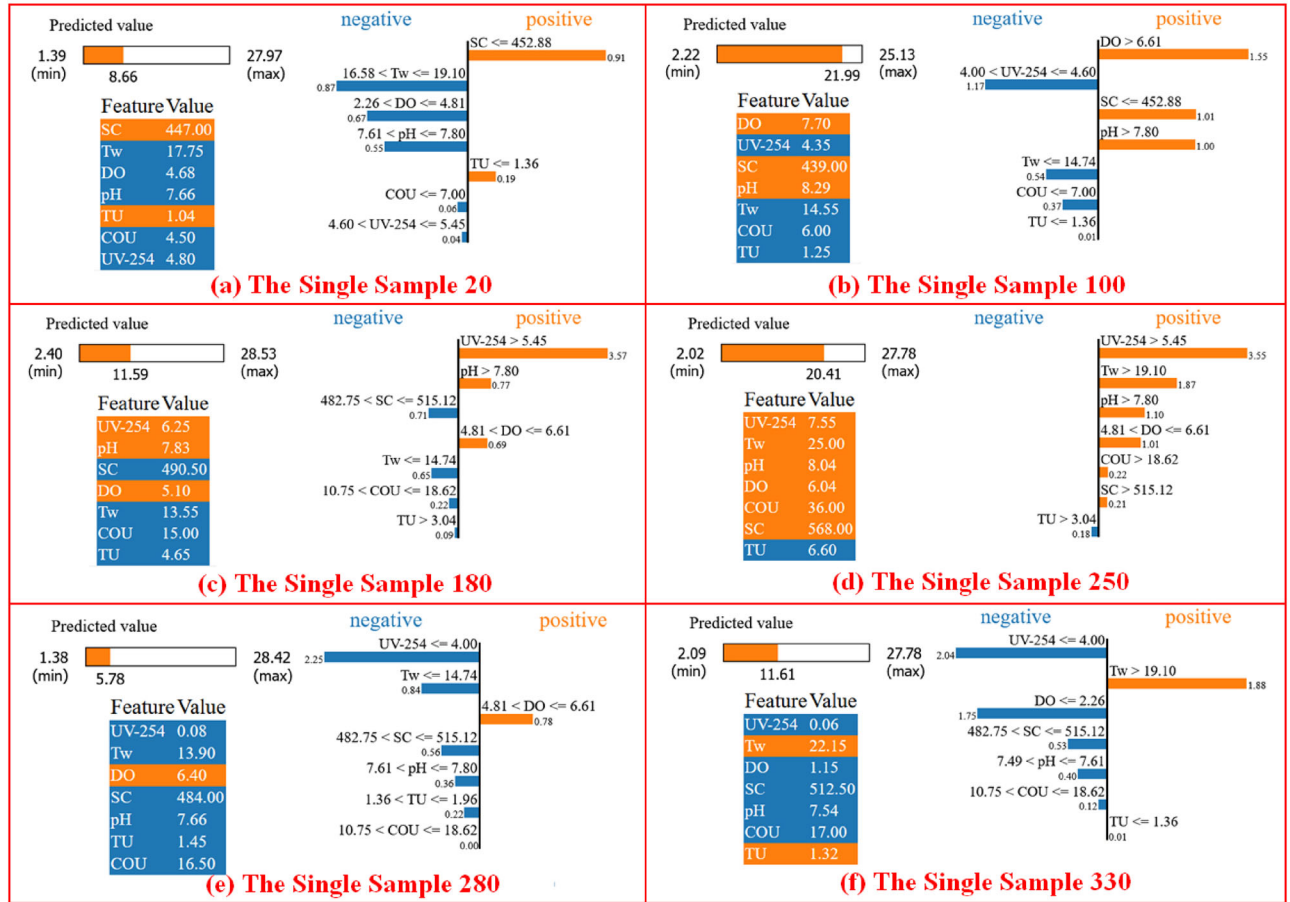


Figure 14. Local LIME bar plots illustrating feature contributions to individual coagulant dosage predictions for six representative samples (a–f).

Several recent contributions published in this journal further contextualize the present findings and clarify its distinct positioning. Y. Zhao et al. (2025) developed multi-source, data-driven machine-learning models for the monitoring of water-quality parameters, thereby underscoring the value of heterogeneous data integration; by contrast, the present study concentrates on a parsimonious set of operationally measured raw-water predictors that directly govern coagulant demand. Gharehbaghi et al. (2025) proposed interpretable intelligent frameworks for the estimation of river-water turbidity, in which SHAP-based attribution was employed in a manner methodologically analogous to the present work, although both the target variable and the operational decision context differ. Li et al. (2024) applied explainable machine-learning models, interpreted through SHAP, to the estimation of daily dissolved-oxygen concentrations in the Tualatin River, and likewise identified specific conductance as a dominant explanatory variable, in agreement with the prominent role of conductivity reported herein. In a related vein, Alizamir et al. (2024) employed SHAP-based explanation for the prediction of

reverse-osmosis desalination-plant performance. Collectively, these studies corroborate the growing convergence towards interpretable ensemble and data-driven modelling for water-treatment applications; the present study advances this line of enquiry by uniting hierarchical feature elimination, dual statistical testing, and combined global-local interpretation within a single coagulant-dosage prediction framework.

The comparative analysis of the applied boosting techniques demonstrated nuanced performance differences. As can be interpreted by the calculated values of statistical measures, CatBoost and GBRT achieved nearly perfect training metrics ($R > 0.999$, $NSE > 0.999$), yet CatBoost demonstrated superior validation performance ($R = 0.922$ versus GBRT's 0.909).

Exploring the inner algorithmic characteristics uncovers the fundamental reasons for the performance differences between the analysed boosting algorithms. Conventional gradient-boosting algorithms such as GBRT, XGBoost, and LightGBM build weak learners in sequence by computing the gradients of the loss function on the same set of training data that was used to construct the

current tree ensemble (Friedman, 2001; Liao et al., 2020; Trizoglou et al., 2021; Y. Zhou et al., 2022). Such structural interdependence leads to an issue of the so-called prediction shift or target leakage, which means that the computed gradients are inevitably biased, yielding great performance on training data but poor validation performance (Prokhorenkova et al., 2018). The CatBoost algorithm addresses this architecture-related limitation using the concept of Ordered Boosting (Prokhorenkova et al., 2018). Ordered boosting uses the random permutation approach to produce unbiased estimates of the gradients for every single data instance in the sample so that the tree is never trained on the gradients obtained from its training set (Prokhorenkova et al., 2018). It is the reason why CatBoost sustains good validation values ($R = 0.922$, $NSE = 0.849$) and avoids the serious overfitting problem that affects GBRT and LightGBM in case of feature matrix calculations. In addition, contrary to leaf-wise (or best-first) tree construction used by LightGBM and XGBoost that can create deep, asymmetric branches which are too sensitive to small noise values (Guo et al., 2023; Osman et al., 2021), CatBoost relies only on symmetric, oblivious decision trees (Prokhorenkova et al., 2018). The idea behind oblivious trees is to use the same split criteria at all nodes of a particular tree layer, which creates a balanced tree topology and acts as a very strong regularizer. This might indicate that the categorical features and ordered boosting mechanism within the CatBoost technique can effectively prevent the overfitting problem and accordingly enhance generalization.

Furthermore, the SHAP analysis identified UV_{254} as the dominant predictor, and its importance is consistent with the established physicochemical mechanisms of coagulation, as discussed in Section 4.3. This agreement between explainable AI and coagulation chemistry strengthens the interpretability and practical reliability of the proposed CatBoost framework.

Among the applied gradient boosting techniques, XGBoost and LightGBM showed comparable performance during the validation phase ($R = 0.916$ and 0.904 , respectively). These results align with the findings reported by Shin and Lee (2024). They identified XGBoost and LightGBM as top performers for coagulant dosing control in Korean water purification plants. However, it should be noted that they claimed that RFR surpassed the boosting methods and was reported as the best predictive model, whereas the current research results favoured CatBoost over RFR. This outcome might be a result of geographic and climatic factors, in that Korean plants typically process more stable surface waters, while Algerian facilities face extreme seasonal variations.

Apart from deterministic and static ensemble models, current global developments have also led to the use

of intelligent Decision Support Systems (DSS) based on reinforcement learning algorithms (Álvarez Díez et al., 2024). Such systems are aimed at dynamic adaptation to raw water conditions and provision of coagulant doses in critical weather conditions, such as rainfall and saline intrusion (Álvarez Díez et al., 2024). Although deep learning and reinforcement learning approaches used internationally can cope with such extreme cases, they still need complex computational setup, incorporation into physical copies of the facility, or huge amounts of data accumulated throughout several years (Álvarez Díez et al., 2024; Lin et al., 2023a). As a result of this benchmarking analysis, our model proves once again its capability to fill this crucial niche; our CatBoost model provides very effective predictions and competitive performance metrics similar to those of complex meta-models and auto-stacking models (W. Zhao et al., 2026) while being sufficiently lightweight to be applied directly into local plants with limited historical data infrastructure. Relative to CatBoost, NGBoost and HistGBRT provided slightly weaker results. In spite of the probabilistic predictions made by these models, they were not as reliable compared with CatBoost and XGBoost, which are deterministic boosting methods.

In terms of input variable sensitivity, the progressive reduction of input variables (from input configuration #1 with seven parameters to configuration #6 with two parameters) revealed model-specific robustness. Nevertheless, NGBoost6 maintained acceptable performance even with reduced inputs in configuration #6 ($R = 0.805$ with only SC and UV_{254}), whereas AdaBoost6 and LightGBM6 performed poorly. Needless to say, MLR6 deteriorated substantially.

5.1. Limitations, constraints, and future research directions

Several methodological limitations and constraints should be acknowledged in this study, as below:

- First, the dataset used here covers only a limited period. Although this matter was not under the control of the authors, it may have resulted in missing some rare or extreme events. Hence, this limitation can affect the models' abilities to generalize to infrequent conditions. For comparison, Lin et al. (2023a) employed ten-year datasets in their graph attention multivariate time series forecasting model. Subsequently, their models could capture long-term climatic patterns that were not available in the present study.
- Second, the results of the Kruskal–Wallis test showed significant differences among the predictions of some of the applied models. This finding indicates that

careful selection of the predictive model can have a substantial impact on operational decisions and outcomes.

- Third, while SHAP analysis provides global interpretability on model predictions, it does not necessarily guarantee that explanations for individual predictions are accurate. SHAP is a powerful method for demonstrating the overall influence of each input variable; however, local interpretability methods, such as LIME, should be applied and further validated for specific cases to ensure reliability and practical relevance.
- Fourth, since the dataset spans only a limited period, the long-term effects of climatic variability and inter-annual trends could not be represented. The acquisition and incorporation of multi-year records is therefore recommended for future work, so that the influence of long-term climate change on raw-water quality and the corresponding coagulant demand may be captured, and so that model generalization to rare and extreme hydroclimatic conditions may be strengthened.
- Fifth, the present framework was trained on synchronous daily measurements, in which the predictors and the coagulant dose correspond to the same time step. Such a configuration reproduces the dose associated with prevailing conditions rather than anticipating its variation. The incorporation of antecedent water-quality observations as lagged predictors—for example, values recorded during the preceding hour or day—is consequently recommended, since it would enable the framework to anticipate impending changes in coagulant demand and thereby to support proactive, rather than merely reactive, dosing control.

Future investigations should integrate the data-driven and ML models with real-time Supervisory Control and Data Acquisition (SCADA) systems. This important suggestion was highlighted by Bachri et al. (2025) in their systematic review of AI optimization for coagulation processes. It is also recommended to incorporate other advanced ML techniques such as reinforcement learning. A practical example includes the advantages of using reinforcement learning as demonstrated by Kadhim et al. (2026) for fish farm effluent treatment. Additionally, hybrid physics-informed ML approaches, which combine coagulation-flocculation kinetics with data-driven components, should be considered for potential accuracy enhancement for unprecedented water quality events.

6. Conclusions

The current study applied multiple ensemble Machine Learning (ML) models, including bagging random forest regression (RFR) and seven boosting models (AdaBoost, CatBoost, GBRT, HistGBRT, LightGBM, NGBoost, and XGBoost) to predict coagulant dosage at the Taksebt water treatment plant in Algeria. For constructing the applied models, seven input parameters—specific conductivity (SC), UV_{254} , dissolved oxygen (DO), temperature (T_w), colour (COU), turbidity (TU), and pH—were considered.

Through changing the point of emphasis from purely empirical matching based on jar tests to a meaningful ensemble intelligence framework, this study produces several clear scientific conclusions:

- (i) Algorithmic superiority via gradient bias correction: The failure in performance of various linear regression algorithms and conventional bagging systems under limited inputs confirms the fact that surface water coagulation is highly nonlinear. The sequential tree-boosting models, especially CatBoost, prove to have the highest stability when generalized into real-world settings. This stability can be achieved due to ordered boosting and symmetric tree structure that prevents target leakage and prediction variance.
- (ii) XAI-Driven operational transparency: Combining global SHAP values with LIME provides an effective means to overcome the problem of deploying deep learning models as black boxes. Unlike deep learning models with hidden layers, which are hard to interpret, the proposed method uses a transparent rule-based threshold logic for verifying predictions.
- (iii) Dominance of dissolved organic and ionic drivers: Consistently, feature sensitivity analysis shows that the UV_{254} and conductivity values drive coagulant demand much more than conventional measures such as turbidity. It is evident that the chemical demand in large-scale plants treating surface waters in the Mediterranean climate regions is dictated mainly by variations in dissolved organics and ionic strength.

Future directions include the physical upgrading of treatment plants through the installation of additional sensors to measure other water quality parameters, as well as the development of other ML models based on multiple input variables. Incorporating climate data from several water treatment plants could further enhance model robustness.

Author contributions

CRediT: **Khaled Merabet**: Conceptualization, Data curation, Investigation, Visualization, Writing – original draft; **Il-Moon Chung**: Funding acquisition, Investigation, Supervision, Writing – original draft; **Amel Fouchal**: Data curation, Formal analysis, Investigation, Writing – original draft; **Francesco Granata**: Investigation, Supervision, Validation, Writing – original draft; **Ozgur Kisi**: Investigation, Supervision, Validation, Writing – original draft; **Sungwon Kim**: Data curation, Formal analysis, Investigation, Visualization, Writing – original draft; **Salim Heddami**: Conceptualization, Investigation, Supervision, Writing – original draft; **Mohammad Zounemat-Kermani**: Formal analysis, Investigation, Validation, Visualization, Writing – original draft

Disclosure statement

No potential conflict of interest was reported by the authors. The authors used ChatGPT (OpenAI, version GPT-4) to support language editing and improve the manuscript's clarity and readability. All scientific content, data analysis, interpretation of results, and conclusions were developed and verified solely by the authors, who take full responsibility for the content of this publication.

Funding

Research for this paper was carried out under the KICT Research Program [project no. 20260156-001, Development of IWRM-Korea Technical Convergence Platform Based on Digital New Deal] funded by the Ministry of Science and ICT.

Consent to publish

All authors have consented to the publication of this manuscript.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Khaled Merabet  <http://orcid.org/0000-0003-0830-4722>
 Il-Moon Chung  <http://orcid.org/0000-0003-0163-7305>
 Amel Fouchal  <http://orcid.org/0009-0002-0158-0388>
 Francesco Granata  <http://orcid.org/0000-0002-2268-6600>
 Ozgur Kisi  <http://orcid.org/0000-0001-7847-5872>
 Sungwon Kim  <http://orcid.org/0000-0002-9371-8884>
 Salim Heddami  <http://orcid.org/0000-0002-8055-8463>
 Mohammad Zounemat-Kermani  <http://orcid.org/0000-0002-1421-8671>

References

- Achite, M., Farzin, S., Elshaboury, N., Valikhan Anaraki, M., Amamra, M., & Toubal, A. K. (2024). Modeling the optimal dosage of coagulants in water treatment plants using various machine learning models. *Environment, Development and Sustainability*, 26(2), 3395–3421. <https://doi.org/10.1007/s10668-022-02835-0>
- Achite, M., Samadianfard, S., Elshaboury, N., & Sharafi, M. (2023). Modeling and optimization of coagulant dosage in water treatment plants using hybridized random forest model with genetic algorithm optimization. *Environment, Development and Sustainability*, 25(10), 11189–11207. <https://doi.org/10.1007/s10668-022-02523-z>
- Agirre-Basurko, E., Ibarra-Berastegi, G., & Madariaga, I. (2006). Regression and multilayer perceptron-based models to forecast hourly O₃ and NO₂ levels in the Bilbao area. *Environmental Modelling & Software*, 21(4), 430–446. <https://doi.org/10.1016/j.envsoft.2004.07.008>
- Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production*, 203, 810–821. <https://doi.org/10.1016/j.jclepro.2018.08.207>
- Ahmed, U., Khan, A. R., Mahmood, A., Rafiq, I., Ghannam, R., & Zoha, A. (2024). Short-term global horizontal irradiance forecasting using weather classified categorical boosting. *Applied Soft Computing*, 155, Article 111441. <https://doi.org/10.1016/j.asoc.2024.111441>
- Alizamir, M., Wang, M., Adnan Ikram, R. M., Kim, S., Ahmed, K. O., & Heddami, S. (2024). Developing an efficient explainable artificial intelligence approach for accurate reverse osmosis desalination plant performance prediction: Application of SHAP analysis. *Engineering Applications of Computational Fluid Mechanics*, 18(1), Article 2422060. <https://doi.org/10.1080/19942060.2024.2422060>
- Álvarez Díez, A., Pena Rois, R., Mocanu, I., Orzan, C., Brebenel, C., Stere, J., Muñíos Landín, S., & Fernández Montenegro, J. M. (2024). Reinforcement learning-based DSS for coagulant and disinfectant dosage selection on drinking water treatment plants. *Water Supply*, 24(1), 86–102. <https://doi.org/10.2166/ws.2023.328>
- An, J., Park, J., Yeon, S., Oh, C., Lee, B., Jung, W., Yun, J., & Oh, H. J. (2025). Machine learning-based prediction of coagulant dosing in drinking water treatment plants using polynomial regression with lasso regularization. *Processes*, 13(12), 3829. <https://doi.org/10.3390/pr13123829>
- An, M., Wang, Z., Bao, X., Tan, Z., & Zhu, Z. (2026). Physics-informed dual-modal decomposition and deep hybrid learning with interpretability for reliable prediction of dissolved oxygen in complex riverine systems. *Journal of Hydrology*, 676, Article 135676. <https://doi.org/10.1016/j.jhydrol.2026.135676>
- Babajide Mustapha, I., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. *Molecules*, 21(8), 983. <https://doi.org/10.3390/molecules21080983>
- Bachri, A., Utomo, W. M., Saon, S., & Laksono, A. B. (2025). Optimization of coagulation process in water treatment plant using artificial intelligence: A systematic literature review and framework. *Journal of Robotics and Control (JRC)*, 6(5), 2471–2487. <https://doi.org/10.18196/jrc.v6i5.27621>
- Boumezbeur, H., Laouacheria, F., Heddami, S., & Djemili, L. (2023). Modelling coagulant dosage in drinking water treatment plant using advanced machine learning model: Hybrid extreme learning machine optimized by Bat algorithm.

- Environmental Science and Pollution Research*, 30(28), 72463–72483. <https://doi.org/10.1007/s11356-023-27224-6>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bressane, A., Goulart, A. P. G., Melo, C. P., Gomes, I. G., Loureiro, A. I. S., Negri, R. G., Moruzzi, R., Reis, A. G. D., Formiga, J. K. S., da Silva, G. H. R., & Thomé, R. F. (2023). A non-hybrid data-driven fuzzy inference system for coagulant dosage in drinking water treatment plant: Machine-learning for accurate real-time prediction. *Water*, 15(6), 1126. <https://doi.org/10.3390/w15061126>
- Chakraborty, A., & Goswami, D. (2017). Prediction of slope stability using multiple linear regression (MLR) and artificial neural network (ANN). *Arabian Journal of Geosciences*, 10(17), 385. <https://doi.org/10.1007/s12517-017-3167-x>
- Chen, S. Z., Feng, D. C., Wang, W. J., & Taciroglu, E. (2022). Probabilistic machine-learning methods for performance prediction of structure and infrastructures through natural gradient boosting. *Journal of Structural Engineering*, 148(8), 04022096. [https://doi.org/10.1061/\(ASCE\)ST.1943-541X.0003401](https://doi.org/10.1061/(ASCE)ST.1943-541X.0003401)
- Dasi, H., Ying, Z., & Yang, B. (2023). Predicting the consumed heating energy at residential buildings using a combination of categorical boosting (CatBoost) and meta heuristics algorithms. *Journal of Building Engineering*, 71, Article 106584. <https://doi.org/10.1016/j.job.2023.106584>
- Ding, C., Wang, Z., Tan, Z., & Zhang, X. (2026). Multi-source heterogeneous data-driven interpretable model based on transformer and Kolmogorov-Arnold networks for stream-flow prediction. *Expert Systems with Applications*, 328, Article 132862. <https://doi.org/10.1016/j.eswa.2026.132862>
- Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020, November). Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning* (pp. 2690–2700). PMLR. <https://doi.org/10.48550/arXiv.1910.03225>
- Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley Additive exPlanations (SHAP). *Case Studies in Construction Materials*, 16, e01059. <https://doi.org/10.1016/j.cscm.2022.e01059>
- Etemadi, S., & Khashei, M. (2021). Etemadi multiple linear regression. *Measurement*, 186, Article 110080. <https://doi.org/10.1016/j.measurement.2021.110080>
- Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management*, 225, Article 105758. <https://doi.org/10.1016/j.agwat.2019.105758>
- Feng, D. C., Liu, Z. T., Wang, X. D., Chen, Y., Chang, J. Q., Wei, D. F., & Jiang, Z. M. (2020). Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. *Construction and Building Materials*, 230, Article 117000. <https://doi.org/10.1016/j.conbuildmat.2019.117000>
- Feng, L., Zhang, Y., Wei, X., Wang, M., Niu, Z., & Wang, C. (2025). Interpretable prediction of coagulant dosage in drinking water treatment plant based on automated machine learning and SHAP method. *Journal of Water Process Engineering*, 75, Article 107925. <https://doi.org/10.1016/j.jwpe.2025.107925>
- Fouchal, A., Merabet, K., Kisi, O., Di Nunno, F., Granata, F., Kim, S., Heddami, S., Zounemat-Kermani, M., & Yaseen, Z. M. (2025). Modeling coagulant dosage in water treatment plant using explainable deep learning model-based SHapley Additive exPlanations (SHAP). *Water Conservation Science and Engineering*, 10(3), 109. <https://doi.org/10.1007/s41101-025-00434-4>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Fu, J. (2025). Cost assessment and strength properties estimation of high-performance concrete using a histogram gradient boosting model integrated with individual and ensemble optimisation algorithm. *International Journal of Pavement Engineering*, 26(1), Article 2450081. <https://doi.org/10.1080/10298436.2025.2450081>
- García, M. V., & Aznarte, J. L. (2020). Shapley Additive exPlanations for NO2 forecasting. *Ecological Informatics*, 56, Article 101039. <https://doi.org/10.1016/j.ecoinf.2019.101039>
- Gharehbaghi, A., Heddami, S., Mehdizadeh, S., & Kim, S. (2025). Development of interpretable intelligent frameworks for estimating river water turbidity. *Engineering Applications of Computational Fluid Mechanics*, 19(1), Article 2511886. <https://doi.org/10.1080/19942060.2025.2511886>
- Gharehbaghi, A., Heddami, S., Mehdizadeh, S., & Kim, S. (2026b). A comparative evaluation of deep learning and machine learning models for river suspended sediment concentration forecasting. *Water Resources Management*, 40(1), 15. <https://doi.org/10.1007/s11269-025-04401-5>
- Gharehbaghi, A., Heddami, S., Mehdizadeh, S., Kim, S., & Chung, I. M. (2026a). Next-generation intelligent framework for pan evaporation prediction: Introducing Chebyshev polynomial-based Kolmogorov-Arnold networks. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-53769-w>
- Guillén-Casla, V., Rosales-Conrado, N., León-González, M. E., Pérez-Arribas, L. V., & Polo-Díez, L. M. (2011). Principal component analysis (PCA) and multiple linear regression (MLR) statistical tools to evaluate the effect of E-beam irradiation on ready-to-eat food. *Journal of Food Composition and Analysis*, 24(3), 456–464. <https://doi.org/10.1016/j.jfca.2010.11.010>
- Guo, J., Yun, S., Meng, Y., He, N., Ye, D., Zhao, Z., Jia, L., & Yang, L. (2023). Prediction of heating and cooling loads based on light gradient boosting machine algorithms. *Building and Environment*, 236, Article 110252. <https://doi.org/10.1016/j.buildenv.2023.110252>
- Guryanov, A. (2019, July). Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 39–50). Springer International Publishing. https://doi.org/10.1007/978-3-030-37334-4_4
- He, X., Luo, J., Li, P., Zuo, G., & Xie, J. (2020). A hybrid model based on variational mode decomposition and gradient boosting regression tree for monthly runoff forecasting. *Water Resources Management*, 34(2), 865–884. <https://doi.org/10.1007/s11269-020-02483-x>

- Hoang, N. D. (2025). Histogram-based gradient boosting machine with SHAP-driven interpretability for predicting intensity of urban heat Island effect. *Discover Civil Engineering*, 2(1), 150. <https://doi.org/10.1007/s44290-025-00301-0>
- Jiang, S., Li, J., Zhang, S., Gu, Q., Lu, C., & Liu, H. (2022). Landslide risk prediction by using GBRT algorithm: Application of artificial intelligence in disaster prevention of energy mining. *Process Safety and Environmental Protection*, 166, 384–392. <https://doi.org/10.1016/j.psep.2022.08.043>
- Kadhim, S. H., Almansoori, A. F., Al-Baldawi, I. A., Abdullah, S. R. S., Abbas, K. F., Imron, M. F., & Kurniawan, S. B. (2026). Optimization of coagulation–flocculation treatment for fish farm effluent using green coagulants and recovery of the produced sludge. *Environments*, 13(2), 88. <https://doi.org/10.3390/environments13020088>
- Kalhormohammadi, M., & Khoramipour, S. (2025). Predictive modeling of coagulant dosing in drilling wastewater treatment using artificial neural networks. *Scientific Reports*, 15(1), 30003. <https://doi.org/10.1038/s41598-025-15155-w>
- Karuppiah, S. P., Rose, J. A., Subasini, C. A., & Vinnarasi, F. S. F. (2026). Enhanced medical image analysis using hybrid Henry gas solubility optimization algorithm with optimized AdaBoost stacked neural networks. *Expert Systems with Applications*, 296, Article 128882. <https://doi.org/10.1016/j.eswa.2025.128882>
- Kavzoglu, T., & Teke, A. (2022). Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal for Science and Engineering*, 47(6), 7367–7385. <https://doi.org/10.1007/s13369-022-06560-8>
- Kim, J., Hua, C., Kim, K., Lin, S., Oh, G., Park, M.-H., & Kang, S. (2024). Optimizing coagulant dosage using deep learning models with large-scale data. *Chemosphere*, 350, Article 140989. <https://doi.org/10.1016/j.chemosphere.2023.140989>
- Kumarakulasinghe, N. B., Blomberg, T., Liu, J., Leao, A. S., & Papapetrou, P. (2020, July). Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 7–12). IEEE. <https://doi.org/10.1109/CBMS49503.2020.00009>
- Li, S., Qasem, S. N., Band, S. S., Ameri, R., Pai, H.-T., & Mehdizadeh, S. (2024). Explainable machine learning models for estimating daily dissolved oxygen concentration of the Tualatin River. *Engineering Applications of Computational Fluid Mechanics*, 18(1), Article 2304094. <https://doi.org/10.1080/19942060.2024.2304094>
- Liao, S., Liu, Z., Liu, B., Cheng, C., Jin, X., & Zhao, Z. (2020). Multistep-ahead daily inflow forecasting using the ERA-Interim reanalysis data set based on gradient-boosting regression trees. *Hydrology and Earth System Sciences*, 24(5), 2343–2363. <https://doi.org/10.5194/hess-24-2343-2020>
- Lin, S., Kim, J., Hua, C., Kang, S., & Park, M.-H. (2023b). Comparing artificial and deep neural network models for prediction of coagulant amount and settled water turbidity: Lessons learned from big data in water treatment operations. *Journal of Water Process Engineering*, 54, Article 103949. <https://doi.org/10.1016/j.jwpe.2023.103949>
- Lin, S., Kim, J., Hua, C., Park, M.-H., & Kang, S. (2023a). Coagulant dosage determination using deep learning-based graph attention multivariate time series forecasting model. *Water Research*, 232, Article 119665. <https://doi.org/10.1016/j.watres.2023.119665>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1705.07874>
- Marvin, G., Grbčić, L., Družeta, S., & Kranjčević, L. (2023). Water distribution network leak localization with histogram-based gradient boosting. *Journal of Hydroinformatics*, 25(3), 663–684. <https://doi.org/10.2166/hydro.2023.102>
- Merabet, K., Di Nunno, F., Granata, F., Kim, S., Adnan, R. M., Heddam, S., Kisi, O., & Zounemat-Kermani, M. (2025). Predicting water quality variables using gradient boosting machine: Global versus local explainability using SHapley Additive exPlanations (SHAP). *Earth Science Informatics*, 18(3), 298. <https://doi.org/10.1007/s12145-025-01796-y>
- Namane, P. I., Letshwenyo, M. W., & Yahya, A. (2025). Evaluation of plant-based coagulants for turbidity removal and coagulant dosage prediction using machine learning. *Environmental Technology*, 46(14), 2570–2585. <https://doi.org/10.1080/09593330.2024.2439183>
- Nhat-Duc, H., & Van-Duc, T. (2023). Comparison of histogram-based gradient boosting classification machine, random forest, and deep convolutional neural network for pavement raveling severity classification. *Automation in Construction*, 148, Article 104767. <https://doi.org/10.1016/j.autcon.2023.104767>
- Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z., & Li, Z. (2021). Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports*, 7, 1246–1255. <https://doi.org/10.1016/j.egyr.2021.02.006>
- Nimon, K. F., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, 16(4), 650–674. <https://doi.org/10.1177/1094428113493929>
- Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using Shapley Additive exPlanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214, Article 106584. <https://doi.org/10.1016/j.cmpb.2021.106584>
- Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, 12(2), 1545–1556. <https://doi.org/10.1016/j.asej.2020.11.011>
- Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., & Costa da Silva, E. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors*, 19(13), 2969. <https://doi.org/10.3390/s19132969>
- Park, H., Joo, C., Lim, J., & Kim, J. (2024). Novel natural gradient boosting-based probabilistic prediction of physical properties for polypropylene-based composite data. *Engineering Applications of Artificial Intelligence*, 135, Article 108864. <https://doi.org/10.1016/j.engappai.2024.108864>
- Peng, L., Lv, S. X., & Wang, L. (2024). Explainable machine learning techniques based on attention gate recurrent unit

- and local interpretable model-agnostic explanations for multivariate wind speed forecasting. *Journal of Forecasting*, 43(6), 2064–2087. <https://doi.org/10.1002/for.3097>
- Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G., & Facchinetti, A. (2023). The importance of interpreting machine learning models for blood glucose prediction in diabetes: An analysis using SHAP. *Scientific Reports*, 13(1), 16865. <https://doi.org/10.1038/s41598-023-44155-x>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31. <https://doi.org/10.48550/arXiv.1706.09516>
- Rahman, T., Momin, M. F., Podder, S. K., Li, H., & Zheng, P. (2025). Optimizing NGBoost with dynamic sequential model-based optimization for predicting UHPC compressive strength on heterogeneous datasets. *Materials Today Communications*, 45, Article 112405. <https://doi.org/10.1016/j.mtcomm.2025.112405>
- Randive, P., Bhagat, M. S., Bhorkar, M. P., Bhagat, R. M., Vinchurkar, S. M., Shelare, S., Sharma, S., Beemkumar, N., Hemalatha, S., Kumar, P., Kedia, A., Massoud, E. E. S., Gupta, D., & Lozanovic, J. (2025). Adaptive optimization of natural coagulants using hybrid machine learning approach for sustainable water treatment. *Scientific Reports*, 15(1), 16096. <https://doi.org/10.1038/s41598-025-96750-9>
- Shams, S. R., Jahani, A., Kalantary, S., Moeinaddini, M., & Khorasani, N. (2021). The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO₂ concentration. *Urban Climate*, 37, Article 100837. <https://doi.org/10.1016/j.uclim.2021.100837>
- Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 56(12), 2353–2360. <https://doi.org/10.1021/acs.jcim.6b00591>
- Shi, H., Wei, A., Zhu, Y., Tang, K., Hu, H., & Li, N. (2024). Accurate and robust ammonia level forecasting of aeration tanks using long short-term memory ensembles: A comparative study of Adaboost and Bagging approaches. *Journal of Environmental Management*, 371, Article 123173. <https://doi.org/10.1016/j.jenvman.2024.123173>
- Shin, D., & Lee, J. Y. (2024). Machine learning regression-based prediction model for the autonomous control of coagulant dosing in smart water purification plants. *Water Practice & Technology*, 19(10), 4036–4048. <https://doi.org/10.2166/wpt.2024.230>
- Stefanovska, E., & Pepelnjak, T. (2025). Optimising predictive accuracy in sheet metal stamping with advanced machine learning: A LightGBM and neural network ensemble approach. *Advanced Engineering Informatics*, 65, Article 103103. <https://doi.org/10.1016/j.aei.2024.103103>
- Tahraoui, H., Belhadj, A. E., Moula, N., Bouranene, S., & Amrane, A. (2021). Optimisation and prediction of the coagulant dose for the elimination of organic micropollutants based on turbidity. *Kemija u Industriji*, 70(11–12), 675–691. <https://doi.org/10.15255/KUI.2021.001>
- Tamim Kashifi, M., & Ahmad, I. (2022). Efficient histogram-based gradient boosting approach for accident severity prediction with multisource data. *Transportation Research Record: Journal of the Transportation Research Board*, 2676(6), 236–258. <https://doi.org/10.1177/03611981221074370>
- Tochio, E. L. L., do Nascimento, B. C., & Lautenschlager, S. R. (2023). Coagulant dosage prediction in the water treatment process. *Water Supply*, 23(9), 3515–3531. <https://doi.org/10.2166/ws.2023.219>
- Tosan, M., Gharib, M., Attar, N., & Maroosi, A. (2025). Enhancing evapotranspiration estimation: A bibliometric and systematic review of hybrid neural networks in water resource management. *Computer Modeling in Engineering & Sciences*, 142(2), 1109–1154. <https://doi.org/10.32604/cmescs.2025.058595>
- Trizoglou, P., Liu, X., & Lin, Z. (2021). Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. *Renewable Energy*, 179, 945–962. <https://doi.org/10.1016/j.renene.2021.07.085>
- Wang, L., Guo, Y., Fan, M., & Li, X. (2022). Wind speed prediction using measurements from neighboring locations and combining the extreme learning machine and the AdaBoost algorithm. *Energy Reports*, 8, 1508–1518. <https://doi.org/10.1016/j.egy.2021.12.062>
- Wang, S., Liu, W., Yang, S., & Huang, H. (2024). An optimized AdaBoost algorithm with atherosclerosis diagnostic applications: Adaptive weight-adjustable boosting. *The Journal of Supercomputing*, 80(9), 13187–13216. <https://doi.org/10.1007/s11227-024-05951-y>
- Wu, G. D., & Lo, S. L. (2008). Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Engineering Applications of Artificial Intelligence*, 21(8), 1189–1195. <https://doi.org/10.1016/j.engappai.2008.03.015>
- Xu, K., Han, Z., Xu, H., & Bin, L. (2023). Rapid prediction model for urban floods based on a light gradient boosting machine approach and hydrological-hydraulic model. *International Journal of Disaster Risk Science*, 14(1), 79–97. <https://doi.org/10.1007/s13753-023-00465-2>
- Xu, S., Wang, S., Yang, Y., Zheng, L., & Xiao, F. (2026). Importance analysis of key parameters for coagulation dosage prediction based on machine learning models. *Process Safety and Environmental Protection*, 205, Article 108152. <https://doi.org/10.1016/j.psep.2025.108152>
- Zafar, M. R., & Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3), 525–541. <https://doi.org/10.3390/make3030027>
- Zhang, C., & Liu, L. (2025). Machine learning prediction model for medical environment comfort based on SHAP and LIME interpretability analysis. *Scientific Reports*, 15(1), 39269. <https://doi.org/10.1038/s41598-025-22972-6>
- Zhang, J., Tylor, N., Zhao, W., Rui, R., & He, C. (2025). Enhancing coagulation prediction in water treatment using a similarity score-based piecewise machine learning model. *Journal of Environmental Engineering*, 151(6), 04025023. [https://doi.org/10.1061/\(JOE\)DU.EEENG-7969](https://doi.org/10.1061/(JOE)DU.EEENG-7969)
- Zhang, K., Achari, G., Li, H., Zargar, A., & Sadiq, R. (2013). Machine learning approaches to predict coagulant dosage in water treatment plants. *International Journal of System Assurance Engineering and Management*, 4(2), 205–214. <https://doi.org/10.1007/s13198-013-0166-5>
- Zhang, S., Liu, M., Xie, M., & Lin, S. (2024). Two-stage short-term wind power probabilistic prediction using natural gradient boosting combined with neural network. *Applied Soft*

- Computing, 159, Article 111669. <https://doi.org/10.1016/j.asoc.2024.111669>
- Zhang, Y., Xie, C., Xue, L., Tao, Y., Yue, G., & Jiang, B. (2021). A post-hoc interpretable ensemble model to feature effect analysis in warfarin dose prediction for Chinese patients. *IEEE Journal of Biomedical and Health Informatics*, 26(2), 840–851. <https://doi.org/10.1109/jbhi.2021.3092170>
- Zhang, Y., Zhao, Z., & Zheng, J. (2020). Catboost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology*, 588, Article 125087. <https://doi.org/10.1016/j.jhydrol.2020.125087>
- Zhao, W., Liu, Y., Wang, L., Li, S., Sun, L., & Shuang, R. (2026). A decision support system for coagulant dosage using explainable NSGA-II optimized stacking framework. *Journal of Water Process Engineering*, 83, Article 109672. <https://doi.org/10.1016/j.jwpe.2026.109672>
- Zhao, Y., Chen, M., He, J., & Ma, Y. (2025). Monitoring water quality parameters using multi-source data-driven machine learning models. *Engineering Applications of Computational Fluid Mechanics*, 19(1), Article 2509658. <https://doi.org/10.1080/19942060.2025.2509658>
- Zhou, Y., Wang, W., Wang, K., & Song, J. (2022). Application of LightGBM algorithm in the initial design of a library in the cold area of China based on comprehensive performance. *Buildings*, 12(9), 1309. <https://doi.org/10.3390/buildings12091309>
- Zhou, Z., Qiu, C., & Zhang, Y. (2023). A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. *Scientific Reports*, 13(1), 22420. <https://doi.org/10.1038/s41598-023-49899-0>
- Zhu, J., Su, Z., Wang, Q., Lan, Z., Chan, F. S. F., Han, Z., Wang, Z., Wong, S. W. F., & Ngan, A. C. F. (2024). Surface quality prediction and quantitative evaluation of process parameter effects for 3D printing with transfer learning-enhanced gradient-boosting decision trees. *Expert Systems with Applications*, 237, Article 121478. <https://doi.org/10.1016/j.eswa.2023.121478>