*FUDMA Journal of Sciences (FJS)*
*ISSN online: 2616-1370*
*ISSN print: 2645 - 2944*
*Vol. 10 No. 17, September, 2026, pp 142 – 155*
*DOI:* https://doi.org/10.33003/fjs-2026-1017-5908

# Reliability-Validated Explainable Machine Learning for Alum Dosage Prediction in Surface-Water Treatment

***Ismail A. Mahmoud, Musa G. Abdullahi and Lurwan Garba**

Faculty of Science Department of Physics Northwest University Kano State, Nigeria.

*Corresponding authors' email: Ismailaminumahmoud27@gmail.com

**Received:** 08 Aug. 2026
**Accepted:** 11 Aug. 2026
**Published:** 22 Sept. 2026

**ABSTRACT**
Optimizing alum dosage remains a major operational challenge in surface-water treatment because coagulant demand is governed by nonlinear physicochemical interactions, episodic source-water disturbances, and non-Gaussian variability. This study developed a reliability-validated and explainable machine-learning framework for predicting alum dosage using operational data from the Tamburawa Water Treatment Plant, Kano, Nigeria. The framework integrated distributional analysis, dependency assessment, stationarity diagnostics, principal component analysis (PCA), ensemble modelling, and SHAP interpretation. Strong distributional irregularity was observed, with Ca, TDS, and EC exhibiting pronounced skewness and kurtosis. Turbidity showed the strongest association with alum dosage ($r = 0.7919$), followed by total hardness ($r = 0.5427$), while stationarity tests confirmed turbidity as a temporally reliable predictor (ADF $p = 0.0051$; PP $p < 0.001$). PCA revealed a compact physicochemical structure, with the first three components explaining 66.6% of total variance. RF-GA achieved the best testing performance ($R = 0.7912$, $R^2 = 0.6202$, MAE = 0.1007, RMSE = 0.1414, NSE = 0.6202), although improvement over RF and RF-PSO was marginal. SHAP confirmed turbidity as the dominant predictor, followed by hardness and pH. In All, the framework supports interpretable, turbidity-prioritized, and more sustainable alum dosing in surface-water treatment.

**Keywords:** Alum Dosage Prediction, Surface-Water Treatment, RF-GA, SHAP, Non-Gaussian Data, Treatment Optimization

## INTRODUCTION

The increasing pressures of urbanization, population growth, changing land use, and climate-induced fluctuations in source-water quality pose significant challenges to surface-water treatment plants striving to provide reliable, efficient, and sustainable drinking water (Cotruvo 2017; Din, Muhammad, and ur Rehman 2023). Among the critical unit processes, coagulation plays a central role by regulating downstream filtration efficiency, reducing organic matter, stabilizing suspended particles, and controlling turbidity. Aluminium sulphate (alum) remains the most widely applied coagulant due to its availability, low cost, and effective destabilizing properties. However, determining the optimal alum dosage remains a persistent operational challenge (Chuiprasert et al. 2026). Under-dosing can lead to poor clarification, high residual turbidity, increased sludge production, and potential aluminium accumulation in treated water, whereas over-dosing increases chemical use, operational costs, and sludge volume (Rahmanian et al. 2015). These challenges are amplified by complex interactions among turbidity, hardness, alkalinity, pH, temperature, ionic strength, dissolved carbon dioxide, calcium, and total dissolved solids, which collectively influence alum hydrolysis, charge neutralization, carbonate buffering, and floc formation. In many municipal water-treatment systems, especially in developing and resource-constrained regions, alum dosing remains largely based on empirical rules, jar-test experience, operator judgment, or fixed heuristics (Sun et al. 2024). While such approaches may suffice under stable water-quality conditions, they become unreliable when source water exhibits episodic disturbances, seasonal variability, or sudden hydrochemical shifts. Fluctuations in turbidity, ionic composition, and buffering capacity can alter alum demand within short operational periods, making average-condition-based strategies insufficient (Krewski, Yokel, and Nieboer 2007; LukmanAliyu, Mukhtar, and Abba 2015).

Over the past decade, statistical modeling, soft computing, and machine-learning algorithms have been increasingly applied to water-quality prediction and treatment-process optimization (García-Ávila et al. 2025). Regression-based approaches offer interpretability but often assume linearity and Gaussian errors, whereas more flexible methods, including random forest, gradient boosting, artificial neural networks, support vector machines, and hybrid optimization algorithms, can capture nonlinear interactions among water-quality variables. Nevertheless, many previous studies prioritize predictive accuracy without systematically validating the statistical reliability of operational input data, which often exhibit skewness, heavy tails, episodic outliers, non-stationarity, heterogeneous variance, and intervention-driven fluctuations (S. Abba et al. 2024; Fullerton, Ceballos, and Walke 2016) Ignoring these properties can produce apparently high-performing models with unstable, non-transferable, or physically misleading variable-importance rankings. Multivariate statistical tools such as correlation analysis and principal component analysis (PCA) have been widely used to identify dependency structures, reduce redundancy, and reveal latent hydrochemical patterns (S. I. Abba, Elkiran, and Nourani 2021; Haj et al. 2025) PCA is particularly valuable in coagulation systems because alum demand is influenced by coupled particulate, ionic, thermal, and carbonate-buffering processes rather than isolated variables. Likewise, explainable artificial intelligence methods, especially SHAP, allow quantification of each input variable's contribution to model predictions (Guo et al. 2024; Khosravi et al. 2025) Despite these advancements, PCA,

machine-learning modeling, and SHAP-based interpretation are often applied separately, without integration into a reliability-validated framework. As a result, there remains limited understanding of whether variables identified as important are statistically stable, physically interpretable, and robust under real non-Gaussian operational conditions.

Despite increasing application of machine learning to coagulation and chemical-dosing problems, an important knowledge gap remains in understanding whether the predictors identified by data-driven models are statistically reliable, physically meaningful, and stable under the heterogeneous conditions encountered in full-scale water-treatment plants. Previous studies have largely emphasized predictive accuracy, whereas considerably less attention has been given to the combined effects of heavy-tailed distributions, temporal instability, predictor dependence, and latent physicochemical interactions on alum-dosage prediction. Consequently, although turbidity is widely regarded as an important coagulation parameter, there remains limited quantitative evidence demonstrating whether its influence persists under non-Gaussian and temporally variable operating conditions or whether alum demand is instead governed by a broader and more diffuse combination of water-quality characteristics.

This study addresses this gap through a reliability-validated and explainable machine-learning framework that integrates distributional diagnostics, dependency analysis, stationarity assessment, principal component analysis (PCA), ensemble learning, and SHAP-based interpretation. Beyond model development, the study seeks to generate new process-level knowledge by determining whether alum dosage is governed by a compact and physically interpretable hierarchy of control variables and by identifying the latent physicochemical modes through which particulate loading, ionic composition, and carbonate-buffering conditions influence coagulant demand. The specific objectives are therefore to characterize the distributional behaviour of operational water-quality variables; assess their dependency structure and temporal reliability; identify dominant latent physicochemical control modes using PCA; compare the predictive generalization of RF, RF-PSO, RF-GA, and XGB models; and interpret the best-performing model using SHAP to distinguish primary from secondary controls on alum dosage. By linking statistical reliability, predictive behaviour, and physicochemical interpretation, the study advances alum-dosage modelling from isolated accuracy-based prediction toward a mechanistically informed and transparent decision-support framework for surface-water treatment.

## MATERIALS AND METHODS

### Data Source

The dataset used in this study was obtained from routine operational monitoring at the TWTP in Kano State, Nigeria (Ismail Aminu Mahmoud et al. 2024). The plant operates under naturally varying hydrochemical conditions, providing a realistic environment for examining alum dosage behavior in a municipal water-treatment setting. The dataset represents operational records collected throughout the year 2022 and includes the most significant physicochemical parameters commonly associated with coagulation performance. The selected input variables include water temperature (WT), turbidity (Turb), pH, electrical conductivity (EC), alkalinity (Alk), total hardness (HARD), calcium (Ca), dissolved carbon dioxide ($CO_2$), and total dissolved solids (TDS), with alum dosage as the target variable. These variables were chosen due to their critical influence on floc formation, carbonate buffering, ionic strength, charge neutralization, and alum hydrolysis. All data were recorded under actual plant operating conditions, rather than controlled laboratory experiments or synthetic simulations, capturing both baseline water characteristics and episodic source-water fluctuations. This real-plant dataset provides a robust framework for subsequent statistical reliability assessment, dimensionality reduction, predictive modeling, and explainable AI-based interpretation of alum dosage behavior, ensuring that the analysis reflects realistic treatment dynamics and operational variability

### Data Cleaning and Preprocessing

Before statistical analysis and machine-learning modelling, the dataset was cleaned and preprocessed to improve consistency, comparability, and numerical stability. The raw operational records were first screened for missing entries, abnormal values, duplicate records, and physically implausible observations. Missing values were treated using interpolation where short gaps occurred within the operational sequence, while extreme values were examined using the interquartile range (IQR) criterion (Ismail A Mahmoud, Wakili, et al. 2025). Values identified as potential outliers were not automatically removed because extreme observations in water-treatment systems may represent genuine events such as turbidity surges, runoff effects, or short-term hydrochemical disturbances. Instead, they were assessed for operational plausibility before adjustment (Jibrin et al. 2024). To avoid scale-induced bias, all input and target variables were normalized before multivariate analysis and model development. Normalization was necessary because the variables were measured in different units and ranges, such as temperature, turbidity, conductivity, hardness, dissolved solids, and alum dosage. The normalization procedure transformed the variables into a comparable numerical scale, improving the stability of PCA, machine-learning training, and SHAP-based interpretation. The normalized value was computed as:

$$X_{normalized-i} = \frac{x_{initial-u-}\ x_{range-min}}{x_{\text{range-max}-}\ x_{range-min}} \quad (1)$$

Where $x_{initial-u}$ present the data to be normalized, $x_{range-min}$ and $x_{\text{range-max}}$ present the minimum and maximum data in the variable range and $x_{normalized-i}$ present the normalized After preprocessing and normalization, the dataset was partitioned into training and testing subsets for predictive modelling. A 70:30 data split was adopted, where 70% of the observations were used for model training, and 30% were reserved for independent testing**.** The training subset was used to calibrate the RF, RF-PSO, RF-GA, and XGB models, while the testing subset was used to evaluate model generalization on unseen data. This partitioning strategy ensured that model performance was assessed beyond calibration accuracy and provided a fair basis for comparing predictive robustness across the selected machine-learning algorithms (See Figure 1).

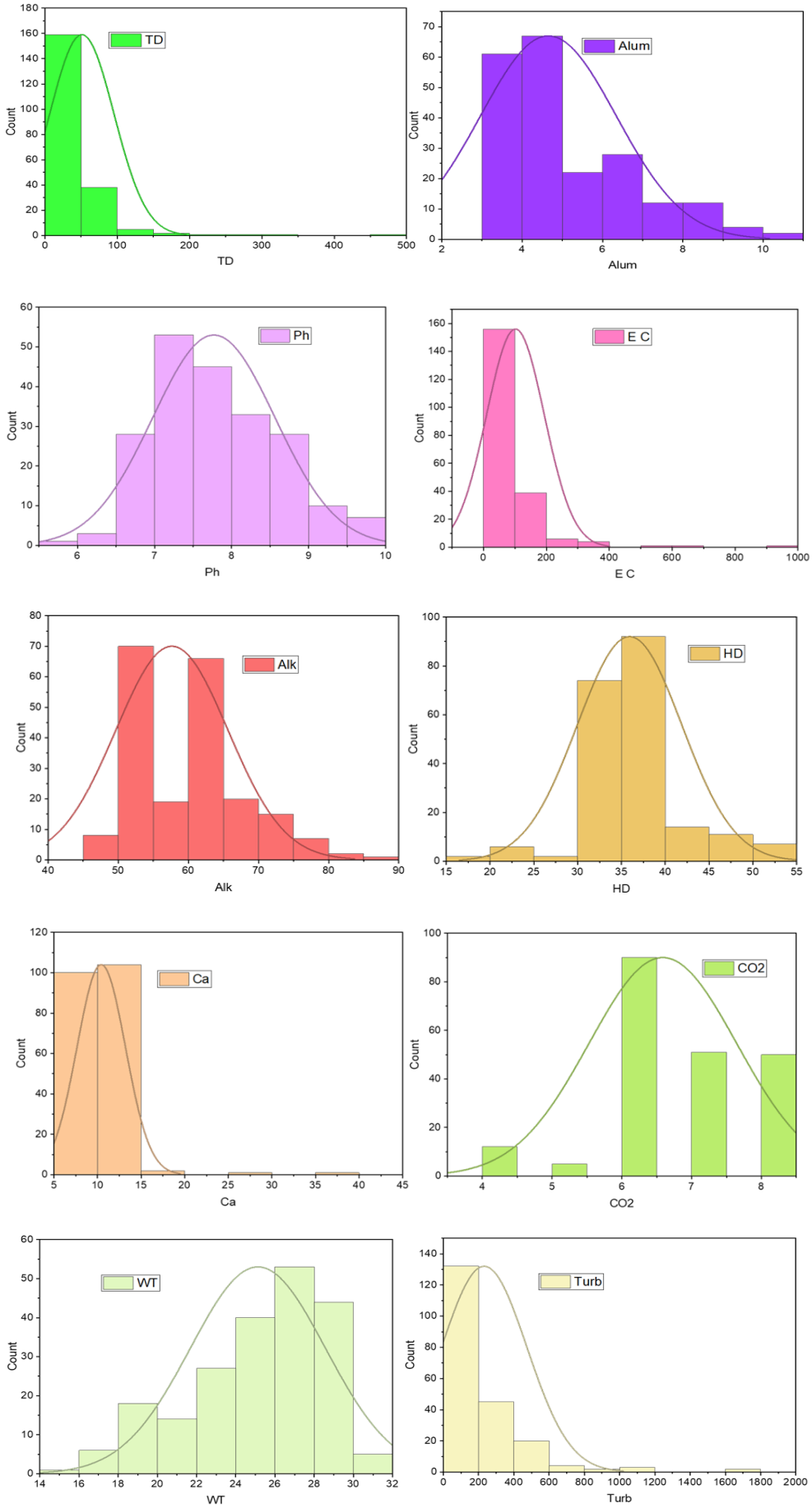


Figure. 1: Basic Data Visualization for Inputs and Target Raw Variables

## MATERIALS AND METHODS

### Model Building

A set of machine-learning models were derived to estimate the alum dosage using the measured physicochemical water-quality variables. The modelling stage aimed at comparing the effectiveness of the stand-alone method with optimised ensemble methods, and at identifying which algorithm could best predict the alum demand in real treatment-plant conditions (Ismail et al. 2025). The models employed were four tree based models: random forest (RF), particle swarm optimization tuned random forest (RF-PSO), genetic algorithm tuned random forest (RF-GA), and extreme gradient boosting (XGB). For the baseline ensemble model, RF was chosen since it consists of several decision trees formed from bootstrapped samples and the final regression output is obtained by taking the average of the tree predictions. This model enables RF to detect nonlinear relationships and interactions among water-quality variables and minimises the instability of each decision tree.

$$YRF\,(x) = \frac{1}{B}\sum_{b=1}^{B} yb(x) \quad (2)$$

where yb (x) is the prediction of the b-th tree.

To improve RF performance, two optimization strategies were introduced. In RF-PSO, particle swarm optimization was used to search for suitable RF hyperparameters by updating candidate solutions according to individual and collective best-performing positions in the search space. In RF-GA, genetic algorithm optimization was used to tune RF hyperparameters through selection, crossover, and mutation processes. These two hybrid models were included to examine whether swarm-based or evolutionary optimization could improve RF generalization for alum dosage prediction.

$$P_i = \frac{Fi}{\sum_{J=i}^{N} fj} \quad (3)$$

$$vi^{(t+1)} = wv^{(t)} + C_1 r_1 (p_i - xi^{(t)} + C_2 r_2 (g - xi^{(t)}) \quad (4)$$

$$xi^{(t+1)} = xi^{(t)} + vi^{(t+1)} \quad (5)$$

Where w is the inertia weight $C_2C_1$ is the cognitive and social coefficients $r_i$ , $r_2$ are random number. XGB was included as a boosting-based comparator. Unlike RF, which builds trees independently, XGB constructs trees sequentially so that each new tree corrects residual errors from previous trees. Its regularized boosting structure gives it strong nonlinear learning capacity, but it may also become sensitive to overfitting when applied to small or heterogeneous operational datasets. Comparing RF, RF-PSO, RF-GA, and XGB therefore provided a balanced assessment of bagging-based, optimized, and boosting-based ensemble learning for alum dosage modelling.

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

Where $L(y_i, \hat{y}_i)$represents the loss function measuring the difference between observed and predicted values, $f_k$ represents the $K^{th}$ regression tree, $K$ is the total number of trees, $\Omega(f_k)$is the regularization term controlling model complexity. The final prediction of XGBoost is expressed as (See Figure 2).

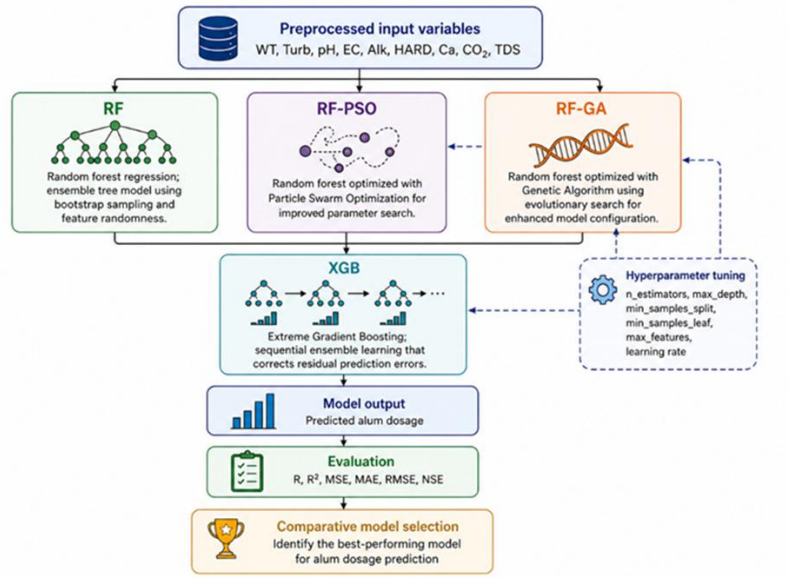


Figure 2: Employed ML Structure of All Models

### Feature Selection via Correlation Analysis

Alum dosage was correlated with water-quality variables as an initial feature-screening step to look at the direction and strength of association. Pearson correlation has been used to analyze whether linear relationships exist and Kendall's Tau has been added to measure the monotonic dependency, even with the presence of non-normal and heavy tailed data. This pairing facilitated both direct linear relationships and rank-order relationships to be assessed (Ismail A Mahmoud, Wakili, et al. 2025). The intent of this stage was not to eliminate variables from the model, but to determine which were the most important physicochemical drivers and to find any redundancy in the predictor variables. This is significant because the dose of alum can be affected by direct parameters (turbidity) as well as modulators (hardness, alkalinity, pH, and parameters related to ionic strength). Thus, based on the above results, correlation analysis was used as a preliminary statistical analysis to gain insight into the interdependence of the variables and to select some variables for further modeling (See Figure 3).

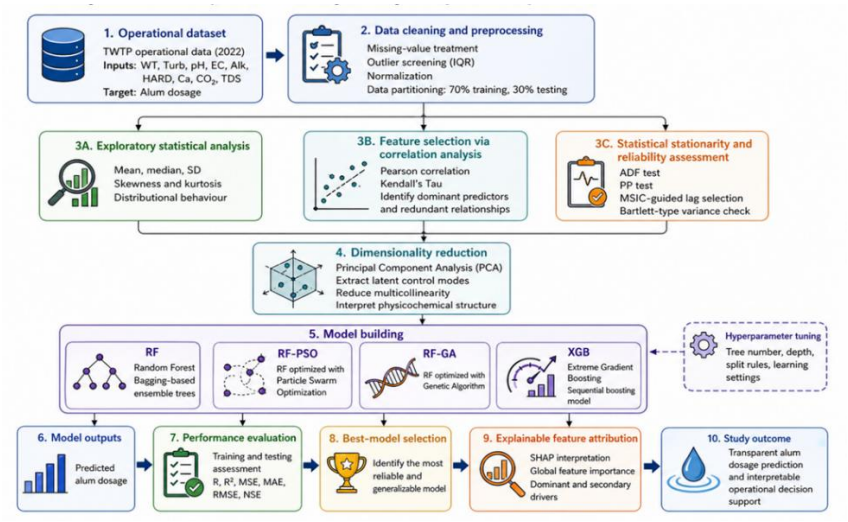


Figure 3: Overview of the Analytical Workflow, Including data Preprocessing, Statistical Reliability Assessment, Dependency Analysis, Dimensionality Reduction, and Explainability-Based Feature Attribution

**Statistical Stationarity and Reliability Assessment**

Stationarity analysis was conducted to evaluate the temporal reliability of the operational water-quality variables. Because treatment-plant records may be affected by seasonal variation, stochastic drift, and short-term hydrochemical disturbances, testing for stationarity helps reduce the risk of drawing conclusions from unstable data patterns. The Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) unit-root tests were applied to determine whether each variable followed a mean-reverting process or exhibited persistent non-stationary behaviour. The ADF and PP tests were used because they provide complementary evidence of temporal stability. The ADF test accounts for serial correlation through lagged difference terms, while the PP test uses non-parametric corrections for serial correlation and heteroscedasticity. Lag selection was guided using the Modified Schwarz Information Criterion (MSIC) to reduce over-parameterization and improve reliability under limited operational observations. In addition, Bartlett-type diagnostics were used to assess variance homogeneity across observational regimes. This reliability assessment ensured that the identified relationships and model inputs were not dominated by short-term fluctuations or stochastic (Gianfreda et al. 2023; Ismail A Mahmoud, Abdullahi, and Garba 2025).

**Mean Stationarity**

A process X(t) is mean stationary if

$$\mu = \mathbb{E}[X(t)] = constant \not\geq t \quad (7)$$

We find the below by computing the sample mean over time and seeing if it changes

$$X = \frac{1}{N}\sum_{T=1}^{N} Xt \quad (8)$$

**Table 1: A Summary Table for Statistical Stationary Analysis**

| Property | Equation | Stationarity Condition |
|---|---|---|
| Mean | $\mu = \mathbb{E}[Xt]$ | Constant over time |
| Variance | $\delta_X^2 = \mathbb{E}[(X(t) - \mu X)2]$ | Constant over time |
| Autocovariance | $\delta^2 = \mathbb{E}[(Xt - \mu)2]$ | Depends only on lag |
| ADF Test | Regression-based test for unit root | Significant negative |

$$X^2 = \frac{(NInS_p^2 - \sum ni\, InS_i^2)}{1+\frac{1}{3(k-1)}(\sum\frac{1}{n-1} - \frac{1}{N-K})} \quad (9)$$

The above equation highlights Bartlett's Test equation, which is used in the context of stationarity testing to assess whether variances are equal across groups. The MSIC, also referred to as the Modified Bayesian Information Criterion (MBIC) is an adaptation of the traditional SIC or BIC**,** developed to improve lag length selection and model reliability in time series analysis, especially under small sample conditions or nonstationary environments.

**Dimensionality Reduction Using Principal Component Analysis**

The PCA was applied to reduce multicollinearity and identify the dominant latent structures within the normalized water-quality dataset. PCA transforms correlated variables into orthogonal components that explain the major patterns of variance in the data. This is useful for alum dosage analysis because coagulation behaviour is influenced by coupled physicochemical processes rather than isolated water-quality parameters (Ismail A Mahmoud, Maiwada, et al. 2025). Component retention was guided by eigenvalue magnitude, scree-plot behaviour, and cumulative explained variance. The retained components were interpreted based on their loading patterns, with emphasis on physically meaningful groupings related to particulate loading, ionic strength, carbonate buffering, and dissolved inorganic chemistry (Kim et al. 2022). In this way, PCA was used not only as a dimensionality-reduction tool, but also as a structural diagnostic for determining whether alum dosage variability is governed by diffuse multivariate influence or by a compact and interpretable control architecture

$$\Sigma = \frac{1}{n-1} Z^T Z \quad (10)$$

S = covariance matrix, and ZT = transpose of the normalized data matrix.

Eigenvalues in PCA

$$\text{Variance explained} = \frac{\lambda_i}{\sum \lambda} \quad (12)$$

**Explainable Feature Attribution**

SHAP-based feature attribution was used to interpret the best-performing predictive model and quantify the contribution of each water-quality variable to alum dosage prediction. SHAP provides an additive explanation of model output by assigning each predictor a contribution value, thereby allowing both global feature importance and observation-level interpretation. Feature importance was calculated using the mean absolute SHAP value across all (Aliyu et al. 2025; Hamilton and Papadopoulos 2023). This enabled the identification of dominant first-order predictors and secondary physicochemical modulators influencing alum dosage. In this study, SHAP was applied after model development so that the interpretation reflected the behaviour of a validated predictive model rather than a standalone statistical ranking. This step provided a transparent link between machine-learning prediction and practical alum dosage control by showing which variables most strongly influenced the model's dosing estimates (Borgonovo, Plischke, and Rabitti 2024).

## RESULTS AND DISCUSSION

**Hyperparameter Tuning**

Hyperparameter tuning was undertaken to enhance the predictive stability of the RF based alum dosage models and to minimise the potential for using weak default model configurations. Two optimization strategies are used here: particle swarm optimization (PSO) and genetic algorithm (GA). Random forest performance is influenced by various parameters like minimum samples for node splitting, maximum tree depth, number of trees, etc. and these parameters are optimized by using these two methods. For the tuning process, an internal validation subset was taken from the training data set and the independent testing data set was reserved for the final evaluation of the models. This configuration minimized data leakage and prevented selection of the model by being exposed to the testing set directly. In the RF-PSO model, PSO was applied for searching the optimal set of the RF hyperparameters. Each particle was a possible solution with the number of trees and maximum tree depth. The ranges for the number of trees searched were from 200 to 500 and for depth (max) from 5 to 20, and the minimum samples at leaves were set to 3. The particles were updated based on the best position of each particle and the best solution of the swarm. Minimization of the validation mean squared error (MSE) was used as the fitness function, which implied that the hyperparameter set chosen was the one that yielded the smallest prediction error on the validation data. In this framework, six particles and six optimization iterations were employed in the PSO search and the acceleration, c1 = 0.5, c2 = 0.3, and w = 0.9 were set. In the RF-GA model, GA was employed as another evolutionary search method for optimizing the hyperparameters of RF. The number of individuals in the GA population corresponded to a possible RF configuration (number of trees and max tree depth). The search ranges were also set between 200 and 500 trees and 5 and 20 for max depth, minimum samples per split was set to 3. Validation MSE was used as fitness function to evaluate candidate solutions. Tournament selection, two-point crossover and uniform integer mutation were employed in the GA search to find better RF configurations for the next generation. The population size was six, crossover probability was 0.5, mutation probability was 0.2, and the number of generations was six. After a number of generations, the best hyperparameter set was selected for the final training of the RF-GA. Also fixed hyperparameters were used in the standalone RF and XGB models for comparative modelling. A 200 tree RF baseline, 12 depth, and 3 minimum split sample was used. XGB was set up with 400 estimators, max depth = 6, learning rate = 0.05, subsample = 0.8, column sample = 0.8. These settings were chosen as a common point for comparison between the optimized RF variants and a conventional RF model and a boosting-based ensemble model. Overall, hyperparameter tuning was performed to enable better generalization of the model instead of just improving the accuracy of the training. It was expected that the RF-PSO model and the RF-GA model would find more suitable tree structures for the alum dosage dataset because of the optimization of the RF configurations accomplished via the internal validation error. It was important for this tuning strategy because overly complex tree structure could be prone to overfitting of operational noise, and overly shallow tree structure could be less able to capture the nonlinear relationship between alum dosage and the physicochemical variables such as turbidity, hardness, pH, etc.

**Exploratory Analysis**

Exploratory statistical analysis of the raw-water system prior to dimensionality reduction and predictive modeling reveals substantial heterogeneity across all physicochemical variables, highlighting the dynamic nature of treatment plant operations. The mean water temperature (WT = 0.604) indicates moderately warm operating conditions, while hardness (HARD = 0.466) reflects a consistent ionic presence in the raw water. In contrast, turbidity (Turb = 0.097), calcium (Ca = 0.145), and total dissolved solids (TDS = 0.074) show lower average values, suggesting episodic fluctuations rather than a sustained baseline. Alum dosage exhibits a mean of 0.233 and a median of 0.143, indicating predominantly low-to-moderate coagulant requirements, with occasional high-demand events. Standard deviations further illustrate system variability: $CO_2$ displays the highest variability (SD = 0.288), followed by WT, pH, alkalinity (Alk), and HARD, reflecting the sensitivity of the treatment process to changes in carbonate balance, thermal conditions, and ionic composition. The alum dosage itself is also highly variable (SD = 0.245), confirming that the plant adapts coagulant dosing dynamically rather than maintaining a fixed regime. The distributions of key inputs are heavily skewed and leptokurtic. EC, TDS, and Ca exhibit extreme positive skewness (5.748–5.936) and kurtosis (43.097–57.444), indicating a concentration of low-to-moderate values with sporadic high-magnitude excursions. Turbidity and pH also display significant skewness (3.388 and -2.231) and high kurtosis (15.282 and 19.261), supporting the presence of episodic operational disturbances. Alum dosage shows moderate right skewness (0.956) and low kurtosis (0.479), suggesting a more controlled response relative to the variability of influent water quality parameters. Negative skew in WT and pH (-1.697 and -2.231) suggests a clustering toward higher normalized readings with fewer low-end extremes. Extreme values are present due to normalization artifacts, with minimum values below zero for several variables (WT, Turb, pH, Alk, HARD, Ca, $CO_2$, TDS, and alum). Maximum values are capped at 1, consistent with min–max scaling but highlighting the influence of operational extremes on normalization. These features emphasize the necessity of explicitly reporting preprocessing methods (e.g., standardization, min–max scaling) and distinguishing between true operational variation and preprocessing artifacts. Collectively, these descriptive statistics reveal that the water-treatment system is characterized by baseline operational stability with intermittent high-intensity fluctuations, both hydrochemical and physical. The observed variability in inputs such as WT,

$CO_2$, and hardness translates into dynamic alum dosing behavior, which is quantitatively reflected in its moderate mean, median, and SD. Furthermore, the combination of high skewness and kurtosis for several inputs underscores the inadequacy of assuming Gaussian distributions in modeling, necessitating robust, nonparametric or machine-learning-based predictive approaches.

**Dependency Analysis of Alum Dosage**

The dependency analysis demonstrates that alum dosage is governed by a hierarchy of operational water-quality variables rather than a diffuse multivariate effect. Using 209 observations, the strongest direct association was observed between turbidity (Turb) and alum dosage, with a Pearson correlation coefficient r = 0.7919, confirming that suspended solids are the principal driver of coagulant requirements. This aligns mechanistically with coagulation theory, as higher particle loads require increased alum for charge neutralization, destabilization, and floc formation. The magnitude of this correlation also quantitatively confirms that plant dosing responds dynamically to particle concentration rather than operating under a static regimen. Hardness (HARD) exhibited the second strongest positive association with alum dosage (r = 0.5427). Although weaker than Turb, HARD's correlation is meaningful: multivalent ions such as calcium and magnesium influence ionic strength and surface charge interactions, modulating coagulation efficiency. This positions HARD as a secondary control variable, modulating alum requirements under variable raw-water chemistry. Other variables including water temperature (WT, r = 0.3399), calcium (Ca, r = 0.2931), and dissolved $CO_2$ (r = 0.2777) show moderate positive correlations, indicating that these parameters provide environmental modulation pathways rather than acting as primary dosing determinants. WT affects hydrolysis kinetics, viscosity, and floc formation; Ca impacts ionic interactions and carbonate chemistry; $CO_2$ contributes to carbonate equilibrium and buffering. Their correlation values quantify their influence as supportive but non-dominant factors in alum dosage control. Alkalinity (Alk, r = 0.0982) and electrical conductivity (EC, r = −0.0141) exhibit negligible direct correlations with alum dosage. Alk's role emerges indirectly through its strong correlation with pH (r = 0.6529), highlighting its function as a buffering variable, stabilizing pH during alum hydrolysis. EC's weak correlation suggests that its influence is embedded within broader ionic structures (TDS, Ca, HARD) rather than expressed in isolation, supporting the need for PCA and multivariate decomposition. In conclusion, the dependency analysis quantitatively establishes the operational prioritization of variables for alum dosage prediction, bridging exploratory diagnostics with PCA-based decomposition and supporting explainable machine-learning models for real-time coagulant optimization.

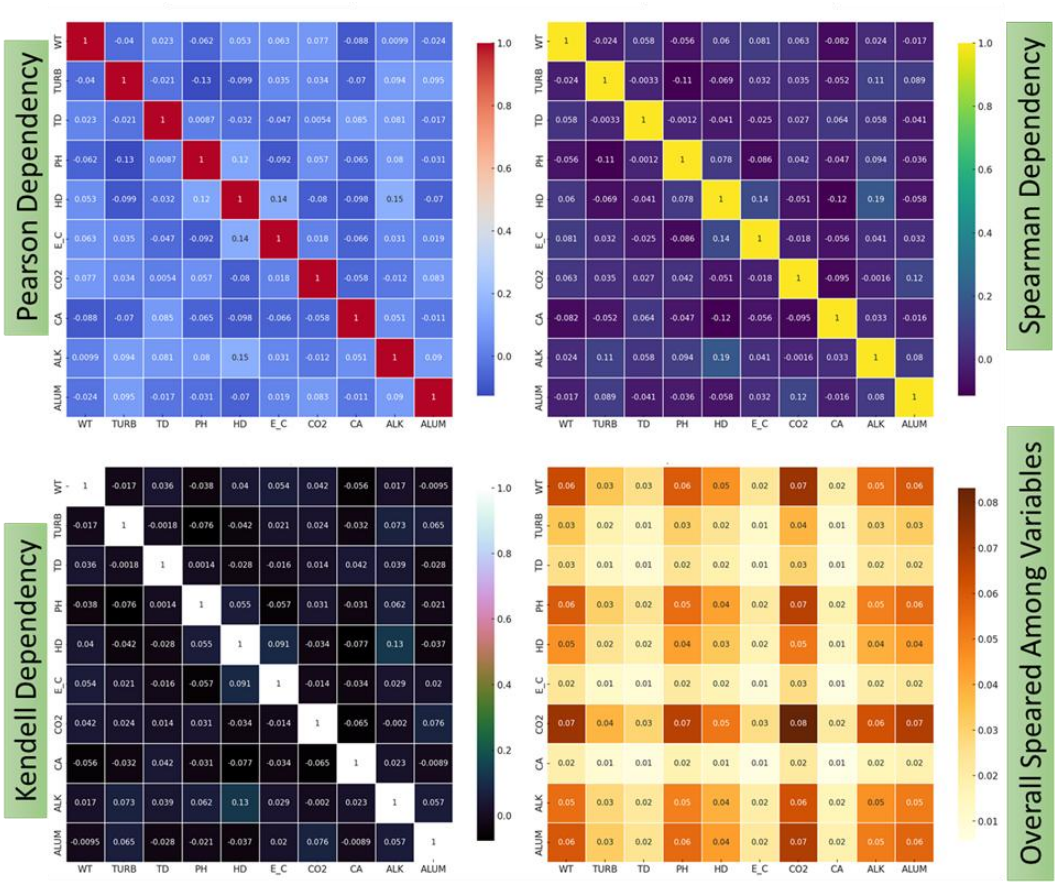


Figure 4: Correlation-Based Dependency Structure Illustrating Dominant Physicochemical Interactions Influencing Alum Dosage Under Real Operational Conditions

**Statistical Reliability and Stationarity Assessment**

To ensure robust structural interpretation and predictive modeling, the temporal stability of operational variables was assessed using Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) unit-root tests. These tests evaluate the null hypothesis of non-stationarity, with rejection at $\alpha = 0.05$ indicating sufficient mean-reverting behavior for reliable modeling. Complementary use of ADF and PP is justified:

ADF accounts for serial correlation through lagged difference terms, while PP corrects for serial correlation and heteroscedasticity non-parametrically. Lag selection employed modified information criteria (MSIC, MAIC) and the Bartlett Kernel (BK) method to mitigate small-sample specification bias. Results indicate that several key operational variables exhibit acceptable stationarity. Water temperature (WT), turbidity (Turb), electrical conductivity (EC), calcium (Ca), dissolved $CO_2$, and alkalinity (Alk) are statistically stationary under at least one testing criterion. Turbidity demonstrates the strongest and most consistent stationarity (ADF p = 0.0051; PP p = 0.0000), reinforcing its reliability as the primary predictor identified in the dependency analysis. This convergence of strong correlation with alum dosage and stationarity confirms that Turb represents a stable, actionable control variable rather than a transient or episodic effect. Several variables display mixed outcomes. Alk is marginal under ADF (p = 0.0772) but strongly stationary under PP (p = 0.0003), suggesting short-term operational fluctuations while maintaining a long-term mean-reverting structure. Similarly, pH meets ADF criteria but fails PP, indicating localized deviations or transient disturbances in acid–base balance. TDS and hardness (HARD) exhibit test-dependent stationarity, reflecting fluctuating dissolved ionic loads influenced by source-water chemistry. Alum dosage shows a complex pattern. It fails ADF stationarity but satisfies the PP–BK criterion, reflecting its role as an operator-controlled response variable. Unlike passive water-quality variables, alum dosage adapts dynamically to fluctuations in turbidity, pH, hardness, and alkalinity. This partial non-stationarity captures adaptive dosing interventions rather than uncontrolled stochastic drift, suggesting that alum dosage is best interpreted as trend-stationary or intervention-sensitive. Quantitatively, the combination of stable environmental predictors (Turb, WT, Ca, $CO_2$) and an adaptive response variable (Alum) supports the validity of predictive modeling frameworks. Stationary predictors provide a reliable basis for feature selection and regression modeling, while trend-sensitive response variables capture operator-driven adaptations, ensuring that predictive models reflect both process dynamics and control interventions. This assessment confirms that the dataset is suitable for PCA, dependency analysis, and machine-learning-based prediction, with turbulence-driven turbidity and its stationarity forming the cornerstone of model reliability.

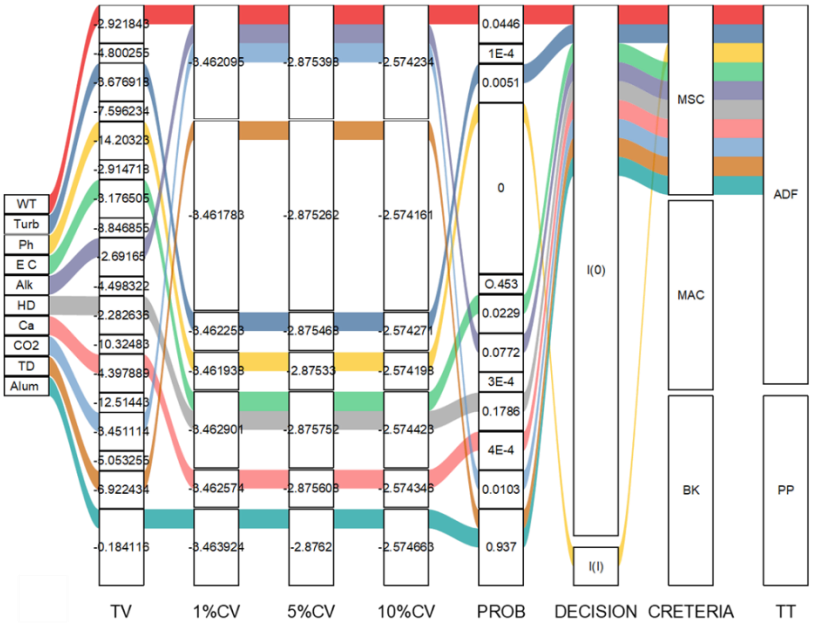


Figure 5: A Comprehensive Reliability Analysis was Carried Out to Evaluate the Stationarity of All the Key Variables

**Dimensionality Reduction Using Principal Component Analysis (PCA)**

Principal component analysis (PCA) was applied to the normalized correlation matrix to identify the latent physicochemical structures governing alum dosage variability and reduce redundancy among correlated operational variables. Coagulation behavior in surface-water treatment is inherently multivariate, influenced by interacting particulate, ionic, thermal, and carbonate-buffering processes rather than isolated factors. Transforming correlated variables into orthogonal components provides a statistically compact representation of the dominant variance structure within the system. The eigenvalue decomposition yielded ten principal components, with the first three explaining a substantial portion of total system variance. PC1 accounted for 29.8%, PC2 for 20.4%, and PC3 for 16.4%, resulting in a cumulative explained variance of 66.6%. When the first five components were retained, cumulative variance rose to 83.6%, indicating that the majority of operational variability is concentrated within a relatively small number of latent dimensions. The scree plot confirmed a sharp reduction in eigenvalue magnitude after the third component, consistent with the Kaiser criterion (first three eigenvalues >1), justifying retention of PC1–PC3 for physical interpretation.

**Table 3: Principal Component Analysis (PCA) Results Showing Eigenvalues, Explained Variance, and Variable Loadings for Water-Quality Parameters Influencing Alum Dosage**

| Number | Value | Difference | Proportion | CV | CP |
|---|---|---|---|---|---|
| 1 | 2.976396 | 0.935092 | 0.2976 | 2.976396 | 0.2976 |
| 2 | 2.041303 | 0.401667 | 0.2041 | 5.017699 | 0.5018 |
| 3 | 1.639636 | 0.750731 | 0.164 | 6.657335 | 0.6657 |
| 4 | 0.888905 | 0.076443 | 0.0889 | 7.54624 | 0.7546 |
| 5 | 0.812462 | 0.30231 | 0.0812 | 8.358702 | 0.8359 |
| 6 | 0.510153 | 0.058546 | 0.051 | 8.868855 | 0.8869 |
| 7 | 0.451607 | 0.118184 | 0.0452 | 9.320461 | 0.932 |
| 8 | 0.333423 | 0.144029 | 0.0333 | 9.653884 | 0.9654 |
| 9 | 0.189394 | 0.032671 | 0.0189 | 9.843278 | 0.9843 |
| 10 | 0.156722 | --- | 0.0157 | 10 | 1 |

The loading structure of PC1 revealed strong positive contributions from alum dosage, turbidity, total hardness, dissolved $CO_2$, and water temperature, representing a particulate–ionic coagulation-demand mode. Turbidity directly drives coagulant requirements through particle destabilization and floc formation, while hardness and $CO_2$ modulate ionic interactions and carbonate chemistry. Temperature further influences reaction kinetics, collectively forming a compact operational control axis. PC2 was characterized by strong positive loadings from total dissolved solids (TDS) and electrical conductivity (EC), representing a dissolved ionic background mode distinct from the particulate-driven response. The moderate negative contribution of pH suggests compensatory acid–base adjustments associated with ionic strength variations. This component captures fluctuations in the chemical environment that indirectly modulate alum demand. PC3 was dominated by pH, alkalinity, EC, and TDS, forming a carbonate-buffering equilibrium mode. Here, the negative loading of alum dosage indicates that under buffering-dominated conditions, alum dosage responds differently compared to the particulate-driven PC1. This is chemically plausible, as alkalinity and pH regulate hydrolysis efficiency and pH stability during coagulation. Higher-order components (PC4 and PC5) contributed smaller proportions of variance, likely reflecting secondary fluctuations from temperature, calcium, and localized operational effects. Their eigenvalues were below unity, indicating they do not represent dominant control modes but may capture seasonal or cycle-specific variability. Collectively, these five components explain more than 83% of total variance, confirming that alum dosage variability is structured and interpretable rather than diffuse or random. These PCA results demonstrate a compressed multivariate structure in the treatment system. The first three principal components summarize the key operational regimes: particulate–ionic coagulation demand, dissolved ionic background, and carbonate-buffering equilibrium. This structured representation validates subsequent predictive modeling and SHAP interpretation, ensuring that the predictor space contains organized, physically meaningful information rather than random multicollinearity.

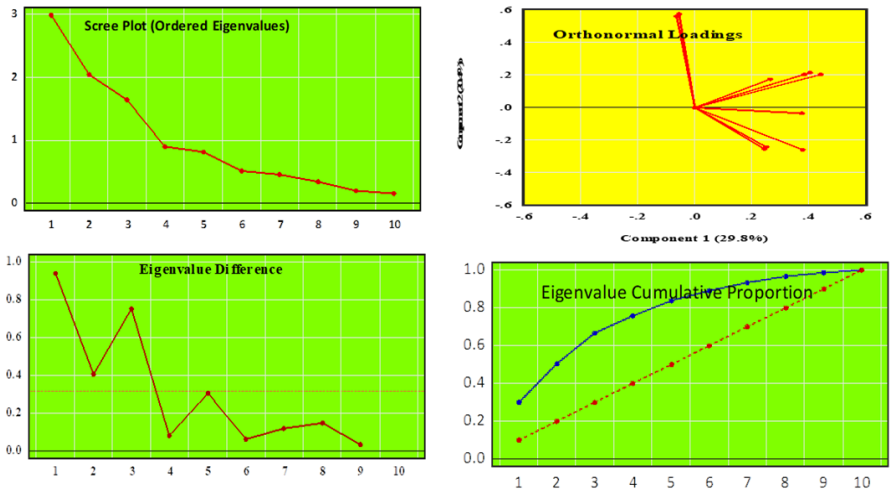


Figure 6: Principal Component Analysis Showing Variance Distribution and Dominant Physicochemical Control Modes Governing Alum Dosage

**Predictive Modelling Performance**

The predictive performance results indicate that all RF-based models achieved comparable testing performance, with RF-GA producing the best but only marginally superior generalization. RF-GA achieved the highest testing $R^2$ and NSE values of 0.6202, together with the lowest testing MAE and RMSE values of 0.1007 and 0.1414, respectively. RF-PSO followed closely, with testing $R^2$ = 0.6195 and RMSE = 0.1415, while the conventional RF model produced testing $R^2$ = 0.6179 and RMSE = 0.1418. The small difference among these three models suggests that RF-based ensemble learning provides stable predictive performance for alum dosage estimation, but the advantage of genetic algorithm optimization is modest rather than substantial. In contrast, XGB achieved near-perfect training performance, with R = 1.0000 and $R^2$ = 1.0000, but its testing performance declined to $R^2$ = 0.5565 and RMSE = 0.1528. This large training–testing gap indicates overfitting, suggesting that the boosting

model captured training-specific patterns that did not generalize as effectively to unseen operational data. Overall, RF-GA was selected as the best-performing model because it provided the strongest testing performance, although its improvement over RF and RF-PSO should be interpreted as marginal (See Figure 7).

**Table 4: Predictive Modelling Results**

| | | RF | RF-PSO | RF-GA | XGB |
|---|---|---|---|---|---|
| Training | R | 0.9833 | 0.9776 | 0.9776 | 1 |
| | $R^2$ | 0.9649 | 0.9538 | 0.9537 | 1 |
| | MSE | 0.0022 | 0.0029 | 0.0029 | 0 |
| | MAE | 0.034 | 0.0418 | 0.0419 | 0.0006 |
| | RMSE | 0.0468 | 0.0538 | 0.0538 | 0.0008 |
| | NSE | 0.9649 | 0.9538 | 0.9537 | 1 |
| Testing | R | 0.7896 | 0.791 | 0.7912 | 0.7461 |
| | $R^2$ | 0.6179 | 0.6195 | 0.6202 | 0.5565 |
| | MSE | 0.0201 | 0.02 | 0.02 | 0.0233 |
| | MAE | 0.101 | 0.1009 | 0.1007 | 0.1099 |
| | RMSE | 0.1418 | 0.1415 | 0.1414 | 0.1528 |
| | NSE | 0.6179 | 0.6195 | 0.6202 | 0.5565 |

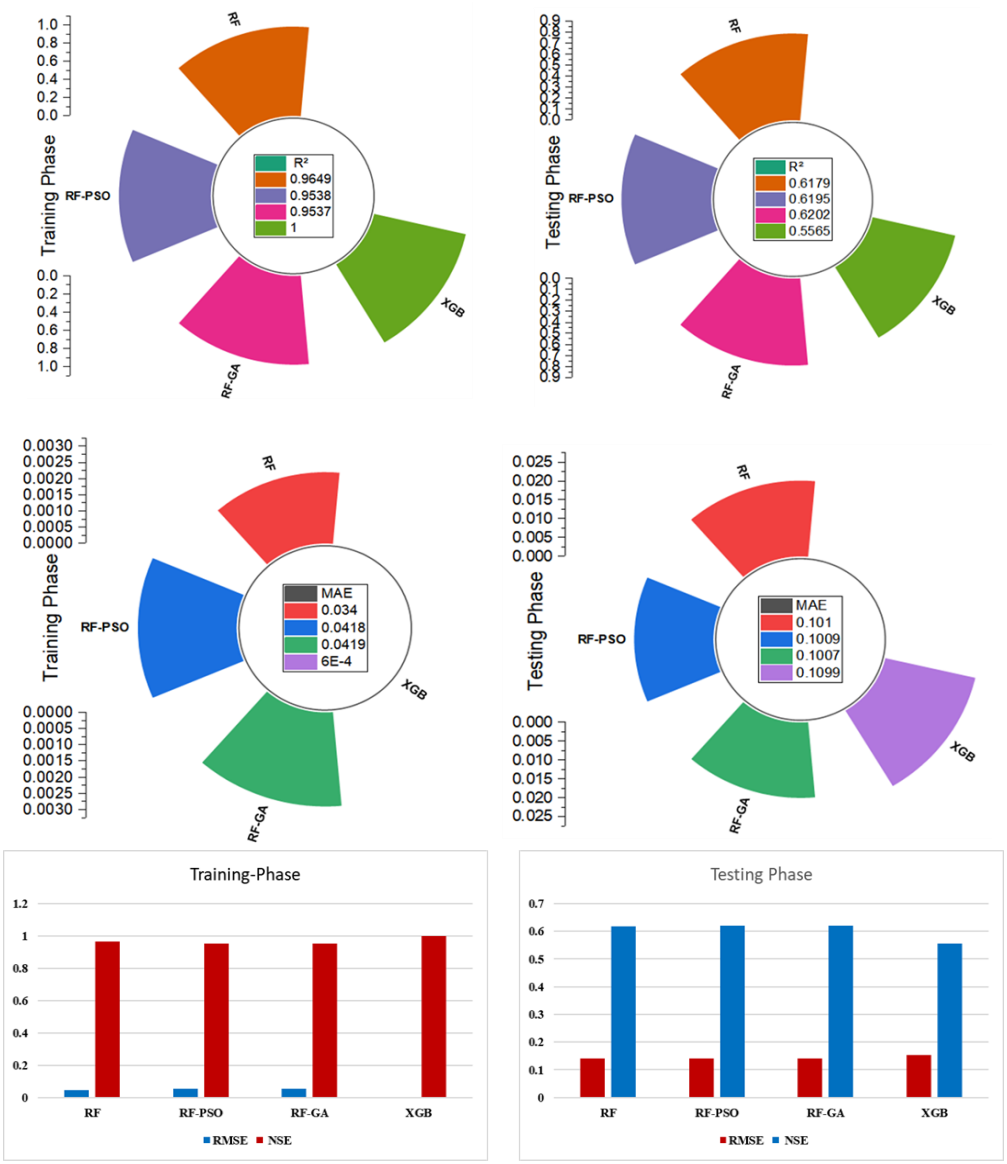


Figure 7: The Performance Metrics Overall Prediction

**SHAP-Based Feature Importance Analysis**

SHAP-based feature attribution was applied to interpret the best-performing predictive model (RF-GA) and quantify the relative contribution of each physicochemical variable to alum dosage prediction. RF-GA was selected because it achieved the strongest generalization performance, with the highest testing $R^2$ and NSE values and the lowest testing errors among the evaluated models. This ensured that feature-

importance interpretation was derived from a robust, reliable predictive structure rather than a poorly generalized model. The SHAP results reveal a highly concentrated feature-importance hierarchy. Turbidity emerged as the dominant predictor, with a mean absolute SHAP value of 0.1732, substantially higher than all other variables. This quantitatively confirms that suspended particle concentration is the primary determinant of alum demand in the treatment system. From a chemical perspective, increased turbidity reflects higher particulate loading, requiring greater alum application for charge neutralization, particle destabilization, and floc formation. This finding corroborates both the correlation analysis and PCA results, where turbidity consistently occupied the main alum-dosing control dimension. Total hardness (HARD) and pH appear as secondary contributors, with mean absolute SHAP values of 0.0108 and 0.0064, respectively. While their contributions are substantially lower than turbidity, they are chemically meaningful. HARD influences coagulation through ionic strength, multivalent ion interactions, and charge-compression effects, which modify particle destabilization and floc growth. Similarly, pH affects alum hydrolysis, aluminium speciation, solubility, and formation of coagulation products. Their lower SHAP magnitudes indicate that they modulate rather than dominate alum demand. The remaining variables including alkalinity (Alk), water temperature (WT), electrical conductivity (EC), total dissolved solids (TDS), dissolved $CO_2$, and calcium (Ca) exhibit small mean absolute SHAP contributions (0.0013–0.0049). Their effects are indirect, interaction-mediated, or conditional, for example, Alk and $CO_2$ affect carbonate buffering and pH stability, while EC and TDS represent the dissolved ionic background. Under operational conditions, these parameters exert less direct influence than turbidity, HARD, and pH. The pronounced disparity between turbidity and the other predictors indicates that alum dosage is controlled by a compact, interpretable hierarchy rather than diffuse multivariate variability. Turbidity governs primary dosing variability, HARD and pH provide secondary chemical adjustment, and remaining variables contribute through background interactions. This confirms that the dominant variable identified in conventional dependency analysis is also the principal driver learned by the nonlinear model. Operationally, these findings support a turbidity-prioritized alum dosing strategy, with real-time turbidity measurements guiding primary adjustments, and HARD and pH serving as auxiliary refinements. This hierarchy enables simpler monitoring, targeted sensor prioritization, and improved chemical-use efficiency in resource-constrained treatment plants. SHAP provides a model-based validation of statistical analyses, linking correlation, PCA, and predictive modeling to practical dosing control. In conclusion, SHAP analysis quantitatively confirms that alum dosage variability is governed by a sharply compressed, physiochemically interpretable feature structure, with turbidity as the principal operational driver, HARD and pH as secondary modulators, and remaining variables contributing primarily through indirect chemical effects. This reinforces the central finding that treatment control is mechanistically structured rather than random, providing actionable guidance for real-time process optimization (See Figure 8).

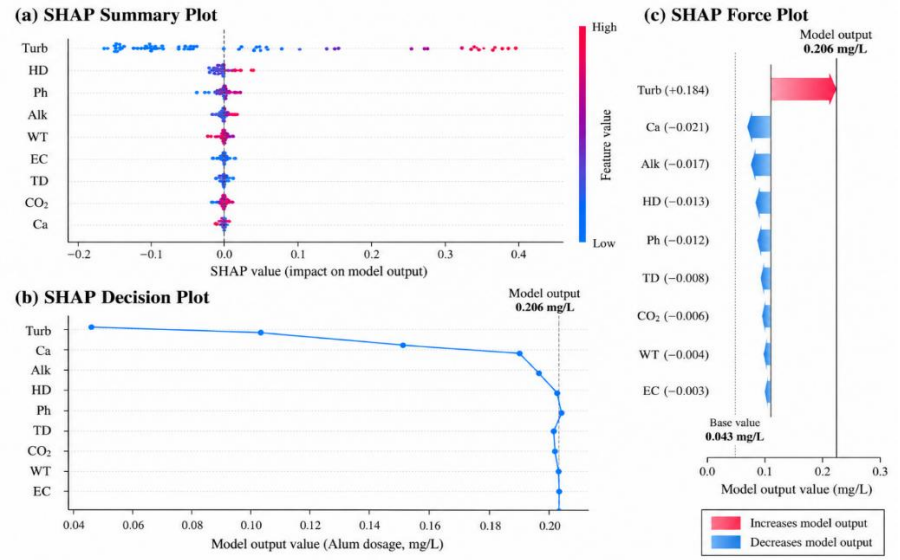


Figure 8: SHAP-Based Global Feature Importance Showing the Dominance Hierarchy of Water-Quality Parameters Influencing Alum Dosage

**Environmental Implications**

The identification of turbidity as the dominant first-order driver of alum dosage has important environmental and operational implications for surface-water treatment systems. Coagulation is one of the most chemically intensive stages of conventional water treatment, and inaccurate alum dosing can affect both treatment efficiency and environmental performance. Excessive alum application increases chemical consumption, sludge generation, handling requirements, and operational cost, while also raising concerns about residual aluminium in treated water. In contrast, insufficient dosing can reduce turbidity removal efficiency and allow particulate-bound contaminants to pass into subsequent treatment stages, particularly during episodic raw-water disturbances and seasonal turbidity surges. The results of this study show that alum dosage is not controlled by a diffuse set of weakly related variables, but by a compact hierarchy dominated by turbidity and secondarily moderated by hardness and pH. This has direct practical value because it supports a simplified and targeted monitoring strategy. Turbidity can serve as the primary operational signal for dosing adjustment, while hardness and pH can provide supporting chemical

information for refining alum requirements under variable ionic and buffering conditions. Such a hierarchy is particularly useful in resource-constrained treatment plants, where continuous monitoring of all possible water-quality variables may not be feasible. From an environmental management perspective, turbidity-prioritized alum dosing can contribute to more efficient chemical use, lower sludge production, reduced sludge-handling energy demand, and improved operational sustainability. Reducing unnecessary alum addition may also decrease the environmental burden associated with chemical production, transportation, storage, and sludge disposal. At the same time, maintaining adequate dosing during high-turbidity events can improve treated-water quality and reduce the risk of particulate breakthrough. Therefore, the proposed framework offers a practical pathway for balancing treatment reliability with chemical-use efficiency. The integration of PCA, dependency analysis, predictive modelling, and SHAP interpretation strengthens the environmental relevance of the findings by showing that the dosing process is physically structured and operationally interpretable. Rather than relying only on empirical adjustment or operator experience, the framework provides data-supported evidence for prioritizing sensor deployment, optimizing coagulant use, and improving process transparency. In this regard, the study supports more sustainable water-treatment operation by linking predictive modelling with chemical management, sludge reduction, and resource-efficient decision-making. These outcomes are consistent with broader sustainability goals related to safe drinking-water provision, efficient resource use, and reduced environmental impact in municipal treatment system (See Figure 9).

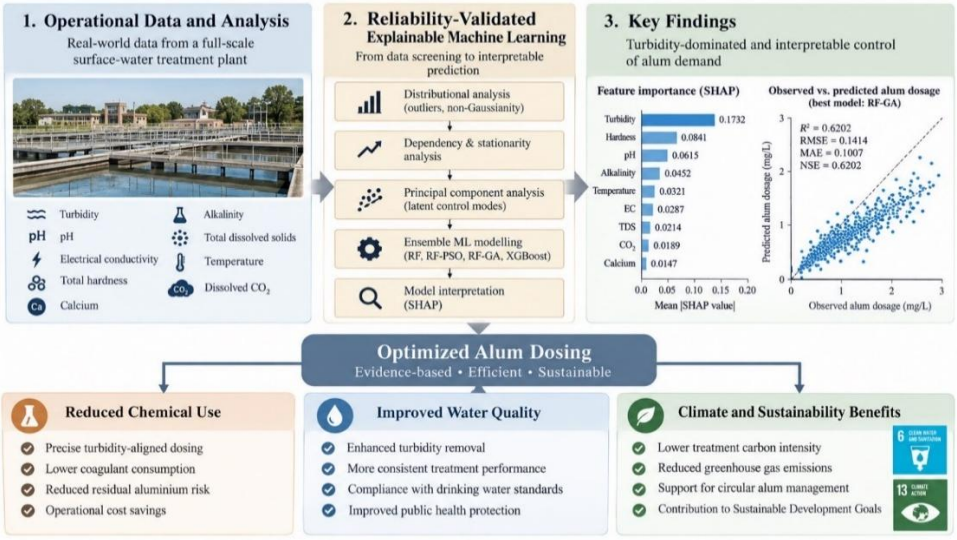


Figure 9: Conceptual Illustration of Environmental and Operational Benefits of Turbidity-Informed, Reliability-Based Alum Dosing Strategies

## CONCLUSION

This study developed an integrated statistical reliability and explainable machine-learning framework for predicting alum dosage under non-Gaussian surface-water treatment conditions at the Tamburawa Water Treatment Plant. Exploratory analysis confirmed that the operational dataset exhibited heterogeneous, skewed, and heavy-tailed distributions, indicating that classical Gaussian assumptions alone are insufficient for robust alum-dosage modeling. Dependency analysis identified turbidity as the dominant physicochemical control variable, with hardness, pH, calcium, dissolved $CO_2$, and water temperature acting as secondary modulators. Stationarity diagnostics confirmed the temporal reliability of turbidity, while principal component analysis revealed a compact control architecture representing three dominant operational modes: particulate–ionic coagulation demand, dissolved ionic background, and carbonate-buffering equilibrium. These results highlight a physically interpretable structure governing alum dosage variability. Machine-learning modeling demonstrated that alum dosage can be predicted reliably from routinely monitored physicochemical variables using tree-based ensemble methods. Optimized random forest models exhibited superior generalization compared to standalone RF and PSO-optimized variants, indicating that evolutionary hyperparameter tuning enhances prediction stability under complex operational conditions. In contrast, boosting-based models displayed signs of overfitting, suggesting that highly flexible algorithms require stronger regularization or larger datasets when applied to treatment-plant records. SHAP-based interpretation of the best-performing model confirmed turbidity as the primary driver of alum dosage, with hardness and pH serving as secondary regulatory variables. Together, statistical analyses, predictive modeling, and explainability assessments demonstrate that alum dosage variability is governed by a hierarchical, mechanistically meaningful structure dominated by particulate loading and moderated by ionic and buffering chemistry, rather than diffuse multivariate randomness. The study is limited by its single-plant scope and specific operational dataset, which may constrain direct generalization to other water-treatment systems and climatic conditions. Key operational variables including flow rate, jar-test outcomes, organic matter, colour, zeta potential, rainfall intensity, mixing conditions, operator decisions, and residual aluminium were not included, potentially contributing to remaining predictive uncertainty. Future research should validate the framework using multi-site and multi-period datasets, integrate additional hydraulic and operational variables, and evaluate real-time deployment for adaptive alum dosing. Further studies should incorporate repeated cross-validation, uncertainty quantification, and multi-objective optimization to minimize alum consumption while maintaining turbidity removal, pH stability, sludge reduction, and treated-water safety.

**DECLARATION OF INTERESTS**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**DATA AVAILABILITY**

The Data Will be Made Available Upon Request

**REFERENCES**

Abba, S. I., Gozen Elkiran, and Vahid Nourani. 2021. "Improving Novel Extreme Learning Machine Using Pca Algorithms for Multi-Parametric Modeling of the Municipal Wastewater Treatment Plant." Desalination and Water Treatment 215(March): 414–26. doi: https://doi.org/10.5004/dwt.2021.26903.

Abba, Sani, Quoc Pham, Anurag Malik, Romulus-Dumitru Costache, Muhammad Gaya, Jazuli Abdullahi, Sagiru Mati, Abdullahi Usman, and Gaurav Saini. 2024. "Optimization of Extreme Learning Machine with Metaheuristic Algorithms for Modelling Water Quality Parameters of Tamburawa Water Treatment Plant in Nigeria." Water Resources Management 39: 1377–1401. doi: https://doi.org/10.1007/s11269-024-04027-z.

Aliyu, Usman U, Ismail A Mahmoud, Sagir Mati, Sukalpaa Chaki, Tukur Abdulkadir, A G Usman, and Sani I Abba. 2025. "Biomass and Bioenergy Optimizing Biomedical Waste Generation Modeling Using Quantum Machine Learning and Economic Development Indicators." 204(July 2025).

Borgonovo, Emanuele, Elmar Plischke, and Giovanni Rabitti. 2024. "The Many Shapley Values for Explainable Artificial Intelligence: A Sensitivity Analysis Perspective." European Journal of Operational Research 318(3): 911–26. doi: https://doi.org/10.1016/j.ejor.2024.06.023.

Chuiprasert, Jedsada, Sutthichai Boonprasop, Krittawit Sopawanit, Tinn Intraluk, Natthapong Taithipmathukon, Thotsatham Takkawatakarn, and Weerawut Chaiwat. 2026. "Design and Process Optimization of an Integrated Turbidity Removal Unit in a Pilot-Scale Continuous-Flow System Using Hybrid Coagulants for Decentralized Water Treatment." Case Studies in Chemical and Environmental Engineering 13: 101327. doi: https://doi.org/10.1016/j.cscee.2026.101327.

Cotruvo, Joseph. 2017. "2017 WHO Guidelines for Drinking Water Quality: First Addendum to the Fourth Edition." Journal - American Water Works Association 109: 44–51. doi: https://doi.org/10.5942/jawwa.2017.109.0087.

Din, Imran Ud, Said Muhammad, and Inayat ur Rehman. 2023. "Groundwater Quality Assessment for Drinking and Irrigation Purposes in the Hangu District, Pakistan." Journal of Food Composition and Analysis 115: 104919. doi: https://doi.org/10.1016/j.jfca.2022.104919.

Fullerton, Thomas M., Alejandro Ceballos, and Adam G. Walke. 2016. "Short-Term Forecasting Analysis for Municipal Water Demand." Journal - American Water Works Association 108(1): E27–38. doi: https://doi.org/10.5942/jawwa.2016.108.0003.

García-Ávila, Fernando, Tatiana Encalada-Alvear, Luis Alvarez-Altamirano, Alex Avilés, Lorgio Valdiviezo, and Emigdio Alfaro. 2025. "Human Health Risk Assessment of Aluminium in Water Treated with Aluminium-Based Coagulants in a Rural Area." Results in Engineering 25: 104200. doi: https://doi.org/10.1016/j.rineng.2025.104200.

Gianfreda, Angelica, Paolo Maranzano, Lucia Parisio, and Matteo Pelagatti. 2023. "Testing for Integration and Cointegration When Time Series Are Observed with Noise." Economic Modelling 125: 106352. doi: https://doi.org/10.1016/j.econmod.2023.106352.

Guo, Zhonghui, Chang Feng, Liu Yang, and Qing Liu. 2024. "Bridging the Gap: An Interpretable Coupled Model (SWAT-ELM-SHAP) for Blue-Green Water Simulation in Data-Scarce Basins." Agricultural Water Management 306: 109157. doi: https://doi.org/10.1016/j.agwat.2024.109157.

Haj, Rachad Al, Mohammad Merheb, Jalal Halwani, and Baghdad Ouddane. 2025. "Baseline Hydro-Geochemical Characteristics of Groundwater in Abu Ali Watershed (Northern Lebanon)." Journal of Hydrology: Regional Studies 57: 102135. doi: https://doi.org/10.1016/j.ejrh.2024.102135.

Hamilton, Robert, and Panagiotis Papadopoulos. 2023. "Using SHAP Values and Machine Learning to Understand Trends in the Transient Stability Limit." IEEE Transactions on Power Systems PP: 1–12. doi: https://doi.org/10.1109/TPWRS.2023.3248941.

Ismail, Ismail I, Mahmud M Jibril, U J Muhammad, Ismail A Mahmoud, Usman U Aliyu, A Abdullahi, and Salim I Malami. 2025. "Ensemble Machine Learning Technique Based on Gaussian Algorithm for Stream Flow Modelling." Techno-computing Journal 1(2): 1–17.

Jibrin, Abdulhayat M, Mohammad Al Suwaiyan, Ali Aldrees, Salisu Dan, Jamilu Usman, Sani I Abba, Mohamed A Yassin, Miklas Scholz, and Saad Sh Sammen. 2024. "Machine Learning Predictive Insight of Water Pollution and Groundwater Quality in the Eastern Province of Saudi Arabia." Scientific Reports: 1–16. doi: https://doi.org/10.1038/s41598-024-70610-4.

Khosravi, Khabat, Aitazaz A Farooque, Amir Naghibi, Salim Heddam, Ahmad Sharafati, Javad Hatamiafkoueieh, and Soroush Abolfathi. 2025. "Enhancing Pan Evaporation Predictions: Accuracy and Uncertainty in Hybrid Machine Learning Models." Ecological Informatics 85: 102933. doi: https://doi.org/10.1016/j.ecoinf.2024.102933.

Kim, Moonil, Park Chul, Wan Kim, and Fenghao Cui. 2022. "Application of Data Smoothing and Principal Component Analysis to Develop a Parameter Ranking System for the Anaerobic Digestion Process." Chemosphere 299: 134444. doi: https://doi.org/10.1016/j.chemosphere.2022.134444.

Krewski, D, Robert Yokel, and Evert Nieboer. 2007. "Human Health Risk Assessment for Aluminum, Aluminum Oxide and Aluminum Hydroxide." J Toxicol Environ Health B Crit Rev 10: 261–69.

LukmanAliyu, L.W Mukhtar, and S.I Abba. 2015. "Evaluation of Coagulation Efficiency of Natural Coagulants ( Moringa Oliefera , Okra ) and Alum , for Yamuna Water Treatment." 2nd International Conference on Science, Technology and Management (January 2015): 2763–71.

Mahmoud, Ismail A, Musa G Abdullahi, and Lurwan Garba. 2025. "Interpretable SHAP-Based Data-Driven Framework for Optimizing Treatment Performance Through Turbidity Dynamics Modelling." 1(3): 16–35.

Mahmoud, Ismail A, Abubakar D Maiwada, Sagir Jibrin Kawu, Mahmud M Jibril, Jamilu Usman, Abdullahi G Usman, and Sani I Abba. 2025. "Multi-Model Environmental Modelling of Energy-Exergy Efficiency Using GUI-Based Aided Design Tools Integrated with Dependency Feature Analysis." Hybrid Advances 10: 100493. doi: https://doi.org/10.1016/j.hybadv.2025.100493.

Mahmoud, Ismail A, Almustapha A Wakili, Abubakar M Danjuma, and A Y Sada. 2025. "Deep Learning LSTM and Random Forest ML-Aided Design Tools for Energy Cooling Capacity Modeling." (March): 1–13.

Mahmoud, Ismail Aminu, Umar Jibrin Muhammad, Sagir Jibrin Kawu, and Mohammed Mukhtar Magaji. 2024. "Machine Learning-Based Wind Speed Estimation for Renewable Energy Optimization in Urban Environments : A Case Study in Kano State , Nigeria." (March). doi: https://doi.org/10.52589/AJSTE-XKYBH2QI.

Rahmanian, Nejat, Siti Ali, Marjan Homayoonfard, N Ali, Mohammad Rehan, Yumna Sadef, and Dr. Abdul-Sattar Nizami. 2015. "Analysis of Physiochemical Parameters to Evaluate the Drinking Water Quality in the State of Perak, Malaysia." Journal of Chemistry 2015: 10. doi: https://doi.org/10.1155/2015/716125.

Sun, Jiasi, Yanran Xu, Haoran Yang, Jia Liu, and Zhen He. 2024. "Machine Learning Facilitated the Conceptual Design of an Alum Dosing System for Phosphorus Removal in a Wastewater Treatment Plant." Chemosphere 351: 141154. doi: https://doi.org/10.1016/j.chemosphere.2024.141154.