Techno-computing Journal 1(3) 16-35 (2025)
https://doi.org/10.71170/tecoj.2025.1.3.16-35

Research Paper

Interpretable SHAP-Based Data-Driven Framework for Optimizing Treatment Performance Through Turbidity Dynamics Modelling

Ismail A. Mahmoud[1*], Musa G. Abdullahi[1], Lurwan Garba[1]

[1]Faculty of Science Department of Physics Northwest University Kano; Nigeria.
Corresponding author: Ismailaminumahmoud27@gmail.com

Abstract

Reliable prediction of Turbidity (Turb) in surface-water treatment systems (WTS) is essential for sustaining safe drinking-water production, particularly in rapidly urbanizing regions where source-water quality (WQ) fluctuates significantly. This study develops a high-precision predictive framework for Turb at the Tamburawa Water Treatment Plant (TWTP) in Kano State, Nigeria, integrating optimized nonlinear models with robust feature-importance diagnostics to improve interpretability and operational usefulness. A comprehensive physicochemical dataset comprising E C (EC), pH, hardness, Alkalinity (Alk), Temperature, alum dosage (Alum), free CO_2, and calcium (Ca) was preprocessed through rigorous screening, normalization, distributional evaluation, and stratified data partitioning. Model development involved the systematic tuning of structural and kernel-based hyperparameters, enabling the construction of four high-performance predictive systems, each with two modelling groups: Neural Network (NN-G1/G2), Bagged Trees (BT-G1/G2), Gaussian Process Regression (GPR-G1/G2), and Support Vector Machine (SVM-G1/G2). Across all configurations, the BT-G1 model delivered the strongest predictive generalization (testing RMSE = 0.0671, MAE = 0.0389), outperforming the NN, GPR, and SVM architectures, and demonstrating high stability across both training and validation phases. SHAP analysis revealed EC as the dominant predictor, followed by free CO_2, while parameters such as Alk and pH contributed comparatively smaller but consistent effects. The findings show that Turb dynamics at TWTP are strongly linked to ionic strength, flow-driven sediment loading, and chemical treatment behavior, aligning with hydrochemical patterns. Beyond model accuracy, the results highlight critical socioeconomic and environmental implications: more precise Turb forecasting can reduce treatment costs, improve allocation of coagulants, and strengthen resilience against climate-driven fluctuations in raw-WQ. The study concludes that interpretable predictive modeling provides a powerful tool for managing WQ risks in northern Nigeria and recommends the integration of real-time monitoring and ensemble-learning extensions in future work to enhance operational decision-making.

Keywords: Turbidity Prediction, Physiochemical Water Quality, Feature Importance Analysis, SHAP Interpretations

1. Introduction

The persistence of degraded WQ in many rapidly growing regions has intensified concerns over public health, ecological stability, and long-term resource security. Among the numerous WQ indicators, Turb remains one of the most immediate and sensitive signatures of contamination, reflecting the presence of suspended solids, colloidal materials, microbial loads, and dissolved organic matter[1]. Elevated Turb compromises treatment efficiency, disrupts aquatic ecosystems, and increases the likelihood of pathogenic survival, making it a central parameter in both regulatory monitoring and scientific

(Received 16 Nov 2025; revised 03 Dec 2025; accepted 10 Dec 2025; first published online 25 Dec 2025); ©Techno-computing Journal 2025. Journal homepage https://technocomputing.org/index.php/tecoj

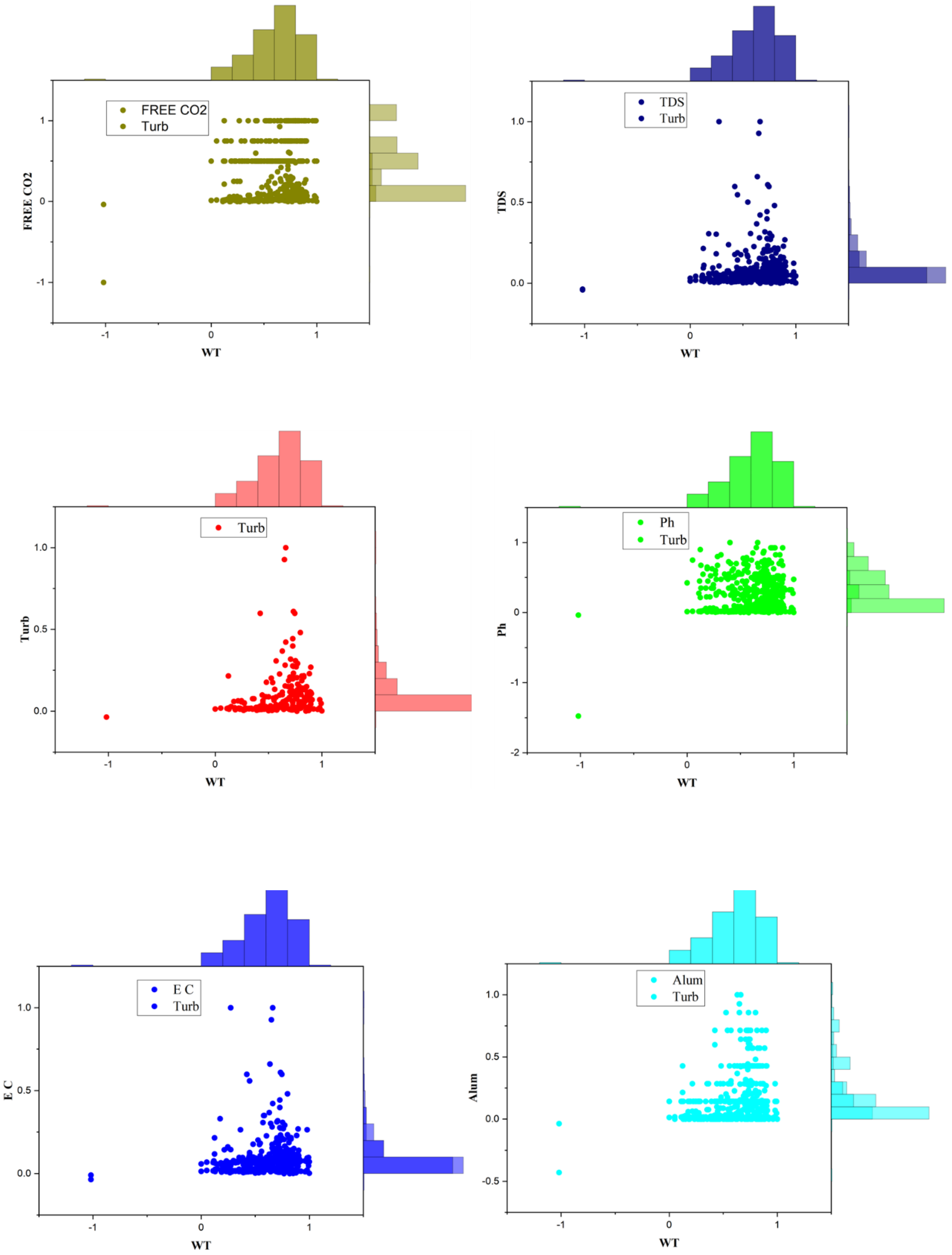
assessments. As global water demand accelerates and urbanization drives changes in land use, the dynamics of Turb have become more complex, necessitating analytical frameworks capable of capturing subtle non-linear interactions among environmental, chemical, and physicochemical variables[2].

Conventional empirical and regression-based methods have provided valuable insights into WQ dynamics, but they often struggle to account for multi-scale relationships embedded within groundwater systems [3]. Recent advances in high-precision sensing, laboratory analytics, and computational modeling have enabled more flexible, non-linear predictive frameworks that outperform traditional approaches in terms of accuracy, generalization, and robustness. Studies have increasingly demonstrated that Turb is influenced not only by visible particulate load but also by deeper hydrochemical characteristics such as E C, hardness, Alk, and Free CO2, which act as proxies for mineral dissolution, surface runoff behavior, and watershed geochemistry (WHO, 2022). These interdependencies highlight the need for modeling strategies that move beyond simple linearity [4].

In response to these challenges, ML methods have gained prominence for modelling complex WQ relationships in treatment plants and distribution systems. NN, tree-based models, and kernel methods have been successfully used to predict Turb, coagulant dosage, and composite WQ indices, often outperforming traditional linear or purely mechanistic approaches when trained on operational data. In particular, previous work at TWTP has demonstrated that data-driven models can predict treated pH, Turb, TDS, and Hard with high accuracy, and that metaheuristically optimized learning algorithms further enhance performance for local WQ parameters. Parallel developments in feature selection and interpretable modelling, such as mRMR, statistical ranking, and SHAP, have shown that it is possible to identify dominant predictors, reduce redundancy, and quantify how individual variables influence model outputs in WQ and related environmental applications[5] [6][7].

Furthermore, global assessments from regions facing rapid population growth and water-demand pressures have emphasized the importance of integrating physicochemical indicators into predictive models to improve risk assessment at early stages [8]. The use of feature-attribution mechanisms such as SHAP values has likewise become increasingly relevant, allowing researchers to identify the relative contributions of each input parameter to model outputs, thereby supporting transparent interpretation and science-based policy planning. By quantifying how each water-quality parameter influences Turb, these methods enhance interpretability, strengthen environmental decision-making, and help align scientific findings with water-management strategies[9][10].

Against this backdrop, the present study develops a Turb prediction framework based on non-linear computational models calibrated using physicochemical variables, including E C, pH, hardness, Alk, Ca, free CO2, temperature, and Alum concentration. The study integrates correlation analysis, statistical significance testing, and detailed SHAP-derived feature importance to clarify each variable's influence on Turb dynamics. By combining high-resolution analytics with interpretable modeling, this research contributes a comprehensive and policy-relevant assessment suitable for regions where Turb remains a persistent environmental and socioeconomic concern. Figure 1 showcases The Raw data instances plot.


FREE CO2
Turb
FREE CO2
WT
1
0
-1
TDS
Turb
TDS
WT
1.0
0.5
0.0
Turb
Turb
WT
Ph
Turb
Ph
WT
-2
E C
Turb
E C
WT
Alum
Turb
Alum
WT
-0.5

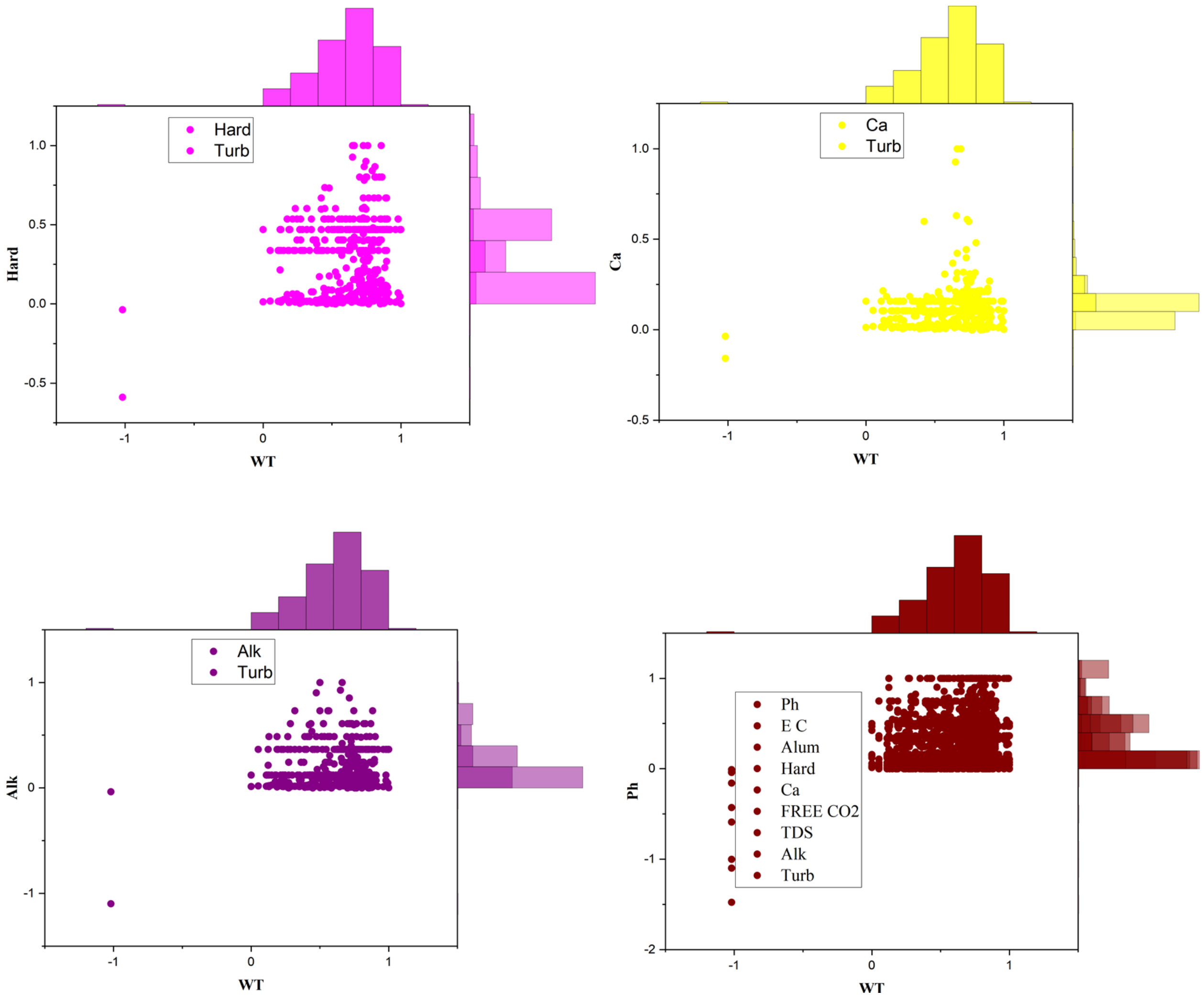


***Figure 1**: Highlights the raw data instances' base plot.*

2.0 Study Location and Data Source

2.1 Study Location

Tamburawa is a prominent town in Dawakin Kudu Local Government, located fifteen kilometers from Kano City in Kano State, Nigeria. The inhabitants of Tamburawa are Hausa, mostly lecturers, farmers, union workers, and businessmen. Irrigation farming is widely practiced. The TWTP is located along the Kano–Zaria Road in Kano State, Nigeria (Latitude: 11.8518° N, Longitude: 8.5359° E) [7]. The plant abstracts raw water from the Challawa River, a tributary of the Kano River, and treats it to supply potable water to Kano metropolis and its surrounding communities [11]. The map of the study area is highlighted in Figure 2. Kano is situated in the semi-arid Sahel region of sub-Saharan Africa, characterized by distinct wet (May–October) and dry (November–April) seasons, which significantly influence raw WQ due to surface runoff, sediment loading, and evaporation rates. These hydrometeorological dynamics make the Tamburawa facility an ideal case study for developing seasonally adaptive coagulant dosing models.

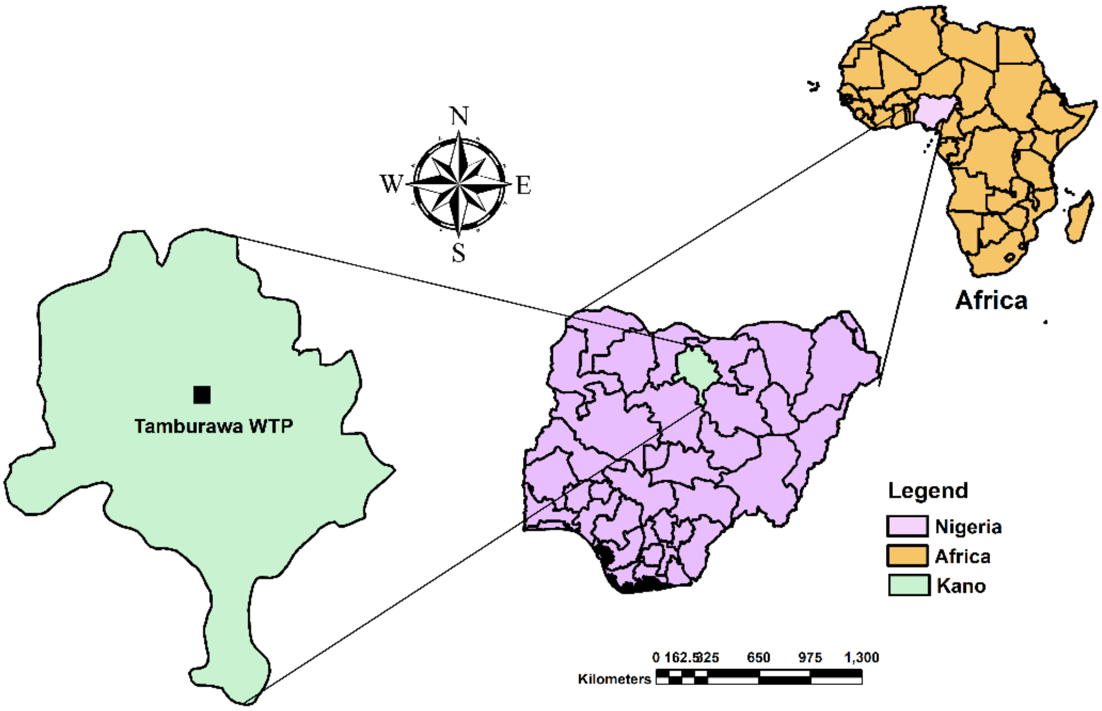


***Figure 2**: Illustrates the map of the study area*

2.2 Data Sources and Preprocessing

This study employed a field-acquired dataset from the TWTP in Kano State, Nigeria. A total of valid observations was collected from daily monitoring logs under real operational conditions during two distinct meteorological periods, dry and wet seasons to evaluate the influence of seasonal variability. Each observation includes measurements of key physicochemical WQ parameters: WT, Turb, pH, EC, Alk, HD, Ca, CO2, and TDS. The response variable is Turb, expressed in mg/L or equivalent normalized units, representing the amount of chemical input required to optimize the process. All measurements adhered to Nigerian Standard for Drinking Water protocols and WHO guidelines, with regular calibration of field instruments by trained plant technicians [13]. An accurate preprocessing pipeline was implemented to ensure data quality and model robustness. Missing values were imputed using linear interpolation to preserve temporal trends. Outliers were identified and removed using the interquartile range (IQR) method, while physically unlikely values were also excluded [14] [15]. All features were normalized using min–max scaling to transform variables into the [0, 1] range, thereby enhancing the stability of gradient-sensitive and distance-based ML models. The normalization followed Equation (1):

$$\mathrm{X}_{\mathrm{normalized-i}} = \frac{\mathrm{x}_{\mathrm{initial-u-}}\ \mathrm{x}_{\mathrm{range-min}}}{\mathrm{x}_{\mathrm{range-max}-}\ \mathrm{x}_{\mathrm{range-min}}} \tag{1}$$

Where $x_{initial-u}$ present the data to be normalized, $x_{range-min}$ a$x_{\mathrm{range-max}}$ present the minimum and maximum data in the variable range and $X_{normalized-i}$ present the normalized data. Seasonal identifiers were retained in the dataset to enable stratified modeling, offering insights into how hydrological and climatic shifts, such as elevated Alum during rainfall or parameter concentration changes in dry seasons, affect Turb. This dual-season structure supports the development of adaptive, season-sensitive predictive models; Furthermore, the dataset was partitioned into two subsets: 70% for model calibration and 30% for validation, ensuring the generalizability of the developed models [16].

3.0 Materials and Methods

3.1 Model Building

The development of the Turb-prediction framework relied on a suite of advanced, data-driven modeling techniques carefully selected to capture the non-linear and multi-scale interactions governing WQ. Each model class was chosen for its demonstrated capacity to extract complex relationships among physicochemical variables, an approach supported by recent hydrological and environmental-data research [17]. The model-building process commenced with the formulation of input–output structures. Turb served as the target variable, while EC, pH, hardness, Alk, calcium, free CO_2, temperature, Alum concentration, and related parameters formed the predictor set. These variables were incorporated based on their known environmental relevance and their statistical contributions identified during preliminary exploratory analysis. The resulting multivariate structure allowed the A diverse array of modeling techniques was employed to improve predictive robustness and avoid methodological bias. NN models (NN-G1 and NN-G2) were constructed to approximate high-dimensional, non-linear relationships. Their architecture consisted of interconnected processing units arranged in layers, enabling the extraction of subtle patterns that are often inaccessible through linear methods, as documented in earlier groundwater-quality studies [18]. Feedforward topology was adopted for its stability and interpretability, with weights optimized during training to minimize error between predicted and observed Turb values. All model families underwent a rigorous cycle of training, validation, and refinement. This involved parameter adjustment, structural optimization, and performance evaluation metrics. Cross-model comparison ensured that no single methodology dominated the analysis, and the final selection of optimal configurations was based strictly on empirical performance rather than algorithmic preference. This multi-framework approach provides resilience against model-specific weaknesses, offering a robust representation of Turb dynamics across varied hydrochemical conditions [19]. Figure 3 showcases the step-by-step modelling process. By integrating multiple modeling strategies, the model-building process established a comprehensive predictive framework capable of capturing the nuanced interactions driving Turb in groundwater. This methodological diversity enhances scientific reliability, supports transparent feature interpretation, and ensures that the resulting predictions align with real-world hydrochemical behaviors. models to represent Turb not as an isolated parameter but as the outcome of interacting chemical and physicochemical processes [20].

3.2 Neural Network (NN)

A NN is a data-driven universal approximator that represents an unknown mapping between an input vector and an output by stacking linear combinations and nonlinear activation functions. Conceptually, the network learns a set of intermediate "features" in a hidden layer and then combines these features to approximate the target function. This provides a flexible way to capture nonlinear relationships and interactions among input variables without specifying a fixed analytical form in advance [5].

3.3 Boosted Regression Tree (BT)

BT has combined many simple decision trees into an additive ensemble that progressively refines the approximation of the target function. Each individual tree partitions the input space into subregions and assigns a constant prediction within each region [21]. On its own, a shallow tree is a weak learner; however, when many such trees are combined through gradient boosting, the ensemble becomes a powerful nonlinear model that can describe interactions and the Threshold effect [22].

3.4 Gaussian Process Regression (GPR)

GPR provides a nonparametric Bayesian framework for regression, where the unknown function f(x) is treated as a sample from a Gaussian process characterized by a mean function and a covariance (kernel) function. Rather than specifying a finite-dimensional parametric form for f(x), GPR defines a distribution over functions and uses the data to update this distribution, yielding predictions with associated uncertainty[23].

3.5 Support Vector Regression (SVR)

SVR seeks a function f(x) that is as "flat" as possible while approximating the training data within a specified tolerance. Flatness is enforced by minimizing the norm of the weight vector in a high-dimensional feature space, while deviations larger than an ε-insensitive band are penalized through slack variables. Nonlinearity is introduced via kernel functions, which implicitly map inputs into the feature space without computing the mapping explicitly [24].

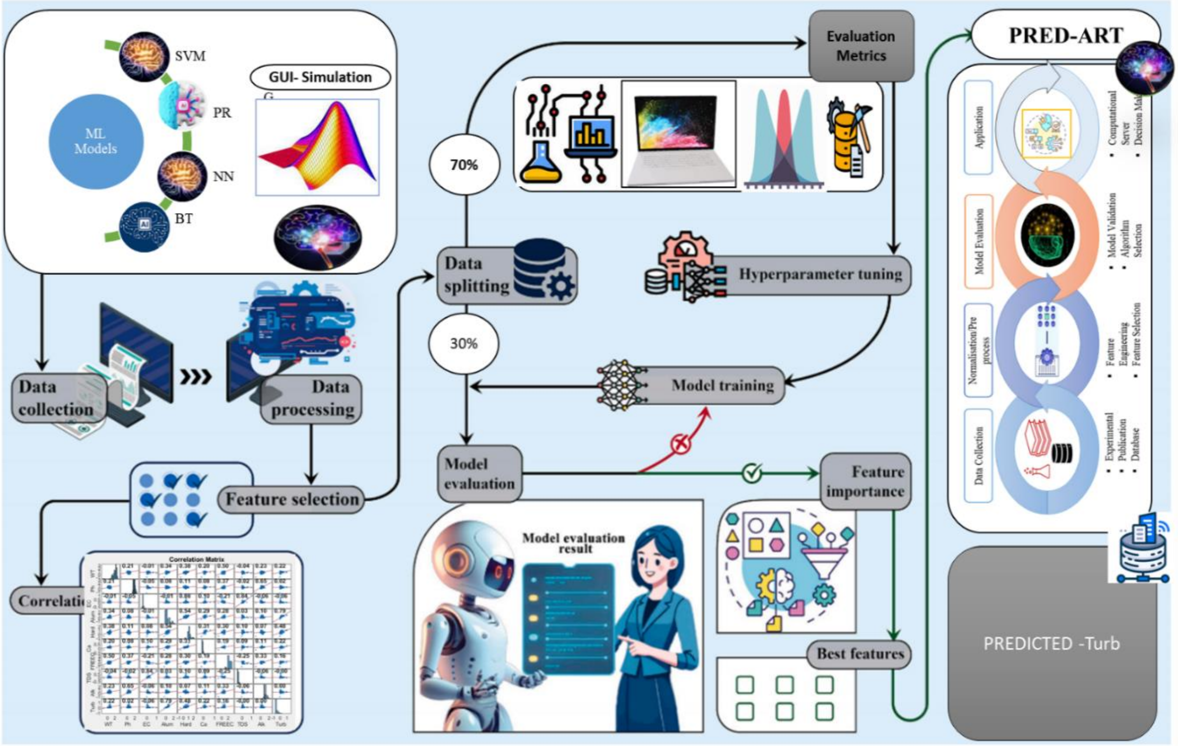


***Figure 3**: The step-by-step modelling Process*

3.6 Feature Selection and Statistical Analysis

Feature selection and statistical analysis were carried out to identify the most informative predictors of Turb among WT, pH, EC, Alum, Hard, Ca, FREECO2, TDS, and Alk, and to reduce redundancy arising from the strong internal correlations within the ionic cluster. Three complementary approaches were used: (i) Pearson correlation analysis to explore pairwise relationships and potential multicollinearity; (ii) mRMR ranking to select variables that are simultaneously relevant to Turb and non-redundant with respect to each other; and (iii) F-test–based ranking to quantify the strength of the linear association between each individual predictor and Turb. In addition, standard error statistics were computed to evaluate the predictive performance of the ML models using the selected features. All computations were performed on the normalized. Pearson correlation analysis was first applied to quantify pairwise linear relationships between Turb and each explanatory variable, as well as among the explanatory variables themselves.

Maximize the average mutual information between the selected features and the target variable

$$D = \frac{1}{|S|} \sum_{\int i \in \mathcal{S}} I(\int ii, \int ic) \quad (2)$$

Minimize the average mutual information among the selected features themselves

$$R = \frac{1}{|S|2} \sum_{\int i \in \mathcal{S}} I(\int ii, \int ij) \quad (3)$$

In addition to the MRMR approach, the F-Test feature selection algorithm was employed to assess the statistical significance of each independent variable with respect to the response variable, alum dosage. The F-test evaluates the ratio of variance between groups to the variance within groups, essentially measuring the discriminatory power of each feature in explaining output variability. [25]. This method is particularly effective in ranking continuous input variables when the target is also constant, as is the case in this study.

$$F = \frac{(\frac{SSR}{P})}{(SSE/(n-p-1))} \quad (4)$$

SSR is the sum of squares due to regression, while SSE is the sum of square of error, p number of predictors and n is the number of observations

However, F-Test is a multivariate, distance-based algorithm that estimates the relevance of features by assessing their ability to differentiate between instances with similar and dissimilar output values [26]. Unlike univariate methods such as the F-Test, which evaluates feature interactions and local instance dependencies, making it particularly effective for nonlinear and correlated datasets.

$$W(A) = \frac{1}{m} \sum_{i=1}^{m} (\frac{|yi-yi'|}{range\ (y)} \cdot \frac{|xi.A-x'i.A|}{range(A)} \quad (5)$$

Where *m* is the number of iterations. xi is a randomly chosen instance. xi’ is the nearest neighbor. yi, y’ is the corresponding measurement of the difference in target values, and (A, xi, x’) is the normalized differences in feature A’s value between xi and xi’

3.7 SHAP-Based Feature Importance Analysis

SHAP were used to quantify the contribution of each input variable to the predicted Turb and to provide an interpretable, model-consistent measure of feature importance. SHAP is based on cooperative game theory and decomposes the prediction of a machine-learning model for a given sample into additive contributions from each input variable. In this study, SHAP analysis was applied to the final selected model (best-performing configuration) trained on the normalized dataset, using the same training-testing split as in the performance evaluation. All SHAP calculations were performed on the normalized features to ensure comparability of contributions across variables [27].

$$\Phi\mathrm{i} = \sum_{S \subsetneq N\{i\}} \frac{S(N)-(S)-1}{(N)} [f(SU\{i\}) - f(S)] \quad (6)$$

Φi where is the value for feature I its contribution to the prediction N: set of all features S; is subset of feature not containing I f(s) is the model output using only feature in subset S f(S u{i}) is the model output using S plus feature I and yet the fraction is a weighting term ensuring fairness over.

3.8 Performance Evaluation Criteria

A determined set of criteria and measurements is used to assess each predictive performance's effectiveness, efficiency, and superiority. They provide a foundation for evaluating performance and helping decision-makers make informed choices about awards, promotions, or performance enhancements. Five statistical metrics were used in this study to assess the models' accuracy: the RMSE, MSE, and MAE. However, Table 1 presents the formal ranges of the performance criteria that are widely used in research to assess the expected performance of the model [14].

***Table 1**: Performance Evaluation Criteria*

Name	Formula	Range
RMSE	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}}(yi - \mathrm{y}^{-})2$	(0 < RMSE < ∞)
MAE	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}/yi - \mathrm{y}^{-}/2$	(0 < MAE < ∞)
MSE	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(yi - \mathrm{y}^{-})2$	(0 < MSE < 100)

The combination of RMSE, MAE, and MSE offers a holistic understanding of model performance from both precision and error perspectives. These metrics were applied consistently across training and testing phases to ensure consistency and detect any potential overfitting.

4.0 Results and Discussion

4.1 Hyperparameter Tuning

Hyperparameter tuning was carried out to ensure that each model operated in a regime that balanced goodness of fit with generalization capability. Rather than relying on default settings, key hyperparameters were systematically varied within plausible ranges, and their combinations were evaluated using the training data, with an internal validation step to assess performance. For all models, the objective was to minimize the prediction error on unseen data, as quantified by the error metrics, while avoiding overly complex configurations that might overfit the training set. For the neural network, tuning focused on the number of hidden neurons, the choice of activation function in the hidden layer, the learning rate, and the maximum number of training iterations [22]. A compact architecture was initially adopted and gradually expanded by increasing the number of hidden neurons until no further reduction in validation error was observed. Learning rates were explored over a range of small values to ensure stable convergence, and early stopping was applied by monitoring the error on a validation subset drawn from the training data. This procedure prevented the network from memorizing noise and ensured that the final configuration provided a smooth yet flexible approximation of the underlying mapping. In the BT model, the main hyperparameters were the number of trees, the maximum depth of each tree, and the learning rate (shrinkage). Shallow trees were preferred as base learners to maintain high bias and low variance at the tree level, while the ensemble progressively reduced bias through boosting. The learning rate was varied over small values, with lower rates generally requiring more trees but yielding more stable generalization. Combinations of tree depth, number of trees, and learning rate were compared using the training–validation split, and the configuration that minimized validation error without evidence of overfitting was selected. For Gaussian process regression, hyperparameter tuning concentrated on the kernel parameters and noise variance. In the radial basis function kernel, the length scale controls how rapidly the function can vary with changes in the input variables, and the signal variance sets the overall amplitude of fluctuations. These parameters, together with the observation noise variance, were adjusted by maximizing the marginal likelihood of the training data, with additional checks based on validation error to avoid excessively small length scales that would produce overly wiggly functions. The final setting represented a compromise between closely tracking the data and maintaining a smooth, physically plausible response surface. In the support vector regression model, the regularization parameter C, the ε-insensitive margin, and the kernel width parameter for the radial basis function were tuned jointly. The parameter C controls the trade-off between model flatness and tolerance of training errors, ε defines the size of the zone within which errors are not penalized, and the kernel width governs the degree of nonlinearity in the mapping. A grid of candidate values for C, ε, and the kernel width was explored, and the combination yielding the lowest validation error while maintaining a stable error pattern between training and validation sets was selected. Across all models, the same training–validation strategy and normalized inputs were used, ensuring that differences in tuned configurations reflected intrinsic model behavior rather than artefacts of the tuning protocol [28].

4.2 Correlation Analysis

The correlation Analysis for WT, pH, EC, Alum, Hard, Ca, FREECO2, TDS, Alk, and Turb shows that the dataset is dominated by a coherent ionic cluster, while Turb behaves more independently. EC is strongly and positively correlated with TDS, Hard, Ca, and Alk, and the corresponding scatter plots show tight, near-linear trends. This indicates that EC and TDS are acting as bulk indicators of dissolved ions, with Hard and Ca representing the contribution of divalent cations, and Alk reflecting the buffering capacity associated with the carbonate system. Together, EC, TDS, Hard, Ca, and Alk form a compact “ionic backbone” that characterizes the chemical regime of the water system. FREE-CO2 also shows moderate associations with EC, TDS, and Hard, suggesting that gas–water interactions and carbonate equilibria are linked to the same geochemical processes controlling the ionic cluster. In disparity, Turb displays weak correlations with EC, TDS, Hard, Ca, and Alk, and its scatter plots against these variables are diffuse. This pattern implies that Turb is largely governed by physical processes such as runoff, erosion, resuspension, and local disturbances, rather than by the gradual changes in ionic strength captured by the chemical variables. WT and pH similarly exhibit only modest correlations with the ionic cluster, indicating that temperature and acid–base conditions are reasonably buffered within the observed range and do not drive large, directly linear changes in Turb. Alum is only weakly correlated with most of the ionic variables, which is consistent with its role as a coagulant rather than an intrinsic component of the natural ionic background. Overall, the correlation structure confirms that while EC, TDS, Hard, Ca, Alk, and FREECO2 describe a strongly coupled chemical subsystem, Turb responds to a partially independent set of drivers. This separation explains why models that combine a well-chosen subset of ionic variables with additional descriptors such as WT, Alum, and FREECO2 are better able to capture the variability in Turb than configurations that treat all variables as equally informative. Yet Figure 4 indicates the Correlation analysis of the variables.

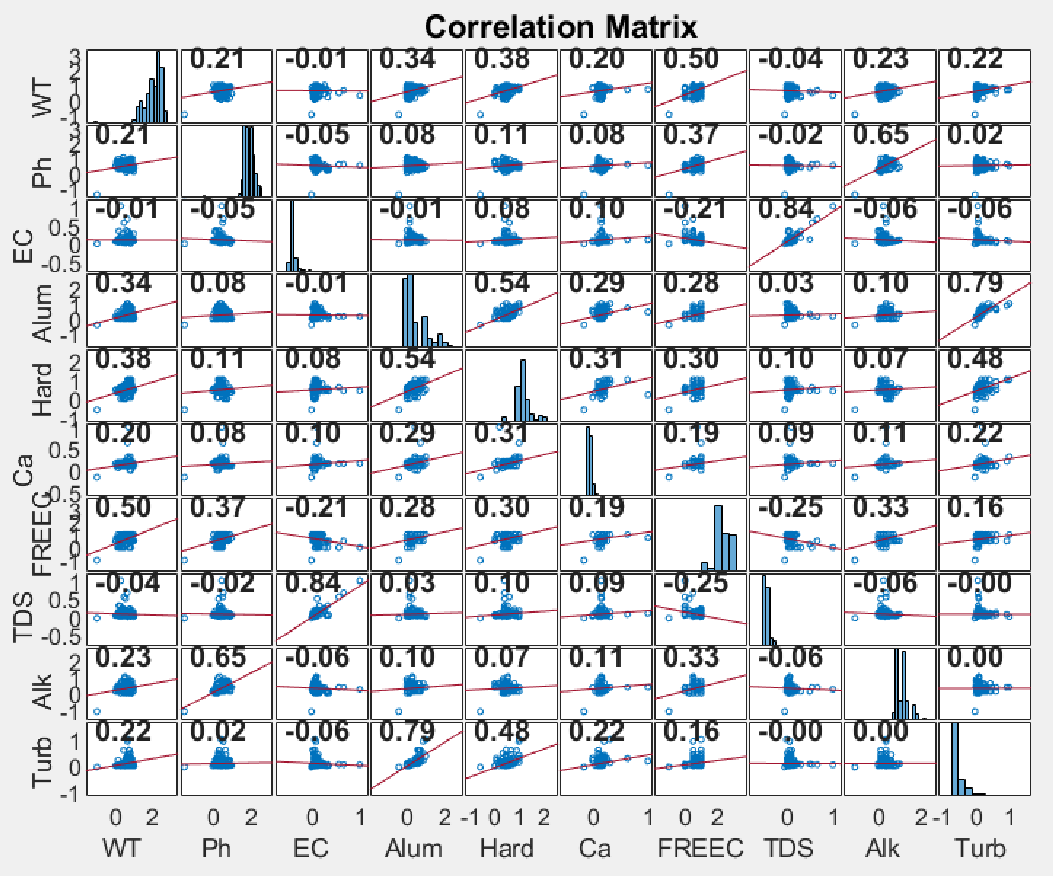


***Figure 4**: Correlation Metrics of variables.*

4.3 Predictive Model Results

Table 2 summarizes the performance of the four employed ML algorithms under the two variable groupings (G1 and G2) using RMSE, MSE, and MAE for both training and testing phases. Across all metrics, the Group 1 configurations clearly outperform the corresponding Group 2 models, confirming the importance of the G1 variable structure for accurate prediction of the target water-quality parameter. Among all configurations, BT-G1 achieves the best overall performance. Its training RMSE (0.0913) and MAE (0.0427) are the lowest in the dataset, and these advantages are retained in the testing phase, where BT-G1 records the smallest RMSE (0.0671) and MAE (0.0389). The similarity between training and testing errors indicates that BT-G1 captures the dominant relationships in the data without overfitting, and that the G1 predictors supply a stable and informative representation of the system. NN-G1, GPR-G1, and SVM-G1 follow closely, with training RMSE values of 0.1205, 0.0973, and 0.0997, and testing RMSE values of 0.0695, 0.0696, and 0.0835, respectively. Their MAE values remain in a narrow band on both datasets, suggesting that all G1 models provide consistent and reliable estimates, even though BT-G1 retains a small but clear advantage.

***Table 2**: The predicted ML Models Results*

	RMSE	MSE	MAE	RMSE	MSE	MAE
NN-G1	0.12053	0.014529	0.055132	0.069471	0.004826	0.040052
NN-G2	0.19404	0.03765	0.119	0.10736	0.011526	0.066624
BT-G1	0.091267	0.00833	0.04273	0.067097	0.004502	0.03891
BT-G2	0.14703	0.021619	0.082982	0.092289	0.008517	0.051825
GPR-G1	0.097334	0.009474	0.048086	0.069628	0.004848	0.07921
GPR-G2	0.13672	0.018692	0.082721	0.11726	0.013751	0.068846
SVM-G1	0.099736	0.009947	0.04716	0.083545	0.00698	0.04176
SVM-G2	0.15004	0.022512	0.084397	0.11293	0.012754	0.45147

The behavior of the Group 2 models contrasts sharply with these results. For each algorithm, the G2 configuration produces larger errors in both phases. NN-G2 shows a substantial increase in training RMSE (0.1940) and MAE (0.1190) compared with NN-G1, and its testing RMSE (0.1074) and MAE (0.0666) also remain noticeably higher. BT-G2 exhibits a similar degradation, with training RMSE rising to 0.1470 and testing RMSE to 0.0923, accompanied by higher MAE values than BT-G1 in both phases. GPR-G2 is particularly affected, with testing RMSE increasing to 0.1173 and MAE to 0.0688, indicating that the additional or altered predictors in G2 introduce variability that the model cannot generalize effectively. The SVM models illustrate the impact of the variable structure most clearly. While SVM-G1 maintains moderate errors (training RMSE 0.0997, testing RMSE 0.0835; testing MAE 0.0418), SVM-G2 displays a drastic deterioration in testing MAE, which jumps to 0.4515 despite a testing RMSE of 0.1129. This unusually large MAE indicates that, under the G2 variable set, the SVM model fails to capture the distribution of the target variable and produces some very large individual prediction errors. This breakdown suggests that G2 likely contains redundant or noisy predictors that disrupt the margin-based decision structure of the SVM, making it highly sensitive to specific observations in the test set. Taken together, these results show that the main driver of predictive performance is not the choice of algorithm alone but the quality and structure of the input variables. Figure 5 highlights a Taylor base plot for MSE for both training and testing phases. All four algorithms benefit from the G1 configuration, which yields low and well-balanced training and testing errors, whereas the G2 configuration consistently degrades accuracy and stability, and in the case of SVM-G2, leads to an almost complete loss of reliability. This confirms that a compact, physically meaningful variable set is essential for building robust machine-learning models for WQ prediction.

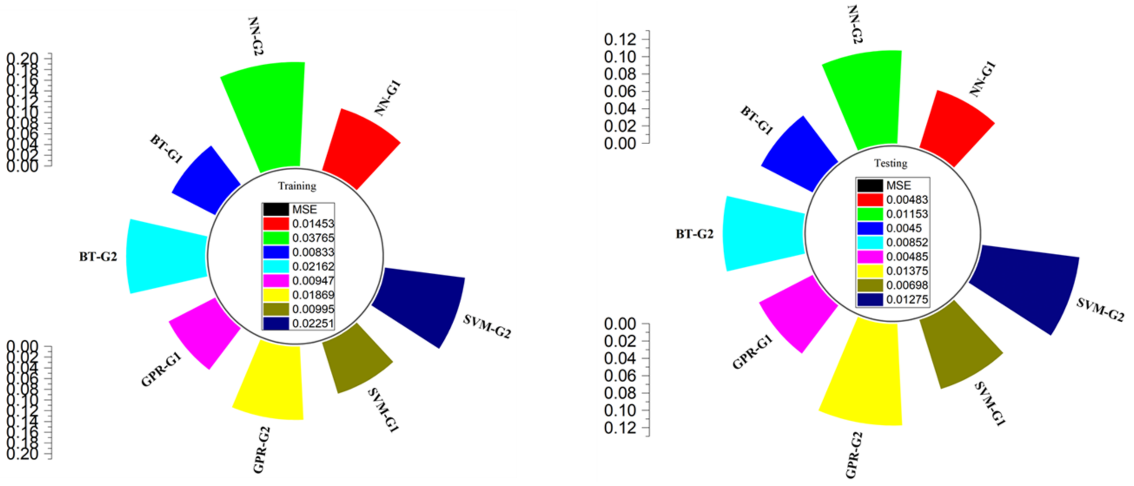


***Figure 5**: Highlighted A Taylor Plot for both training and testing MSE*

4.4 Feature Importance Using the mRMR Algorithm

The mRMR algorithm ranks features by identifying those that provide the highest predictive relevance to the target variable, in this case, Turb, while minimizing redundancy among the predictor variables. The objective is to select a set of features that collectively offer the greatest explanatory power without duplicating information, which is essential in WQ modeling where many physicochemical parameters exhibit high intercorrelation. In the provided mRMR output, Alum emerged as the most influential variable, ranking highest in importance. This observation is consistent with the treatment processes at the TWTP, where alum is the primary coagulant applied to destabilize colloidal particles. Higher alum dosing typically corresponds to raw water of elevated Turb, and its relationship with Turb tends to be monotonic and strong, making the parameter both relevant and non-redundant. The high mRMR score, therefore, reflects its direct operational linkage and sensitivity to Turb fluctuations in surface water systems. The second-ranking feature, TDS, indicates a strong association between dissolved ionic content and suspended matter behavior. Elevated TDS often coincides with anthropogenic runoff or seasonal hydrological inputs that simultaneously raise particulate loads. Although TDS and Turb are not mechanistically identical, their shared environmental drivers explain why the mRMR algorithm retains TDS as highly relevant yet not heavily overlapping with alum or other variables. Hardness and E C follow in the importance hierarchy. Their significance suggests that mineral content and ionic strength exhibit indirect relationships with Turb, possibly through geogenic contributions or water–sediment interactions. EC, in particular, is widely recognized as a key indicator of overall WQ variability in semi-arid regions and contributes unique information that mRMR deems valuable without excessive redundancy. Ca, FREE CO_2, pH, and WT appear in the mid-to-lower ranks. These variables influence floc formation, solubility equilibria, and biological activity; however, their predictive contribution to Turb is comparatively weaker or partly redundant with other parameters. For example, pH affects coagulation efficiency, but its impact is largely mediated through alum dosing and Alk, which may explain why mRMR reduces its importance when alum is already selected. The lowest-ranking variables, such as Alk and Turb itself (when used redundantly in multi-step feature selection), show either redundancy with stronger predictors or limited independent explanatory power. Their lower rank does not imply irrelevance scientifically but highlights that they add minimal additional information to the model once higher-ranking variables are considered, yet ranking instances are highlighted in Figure 6.

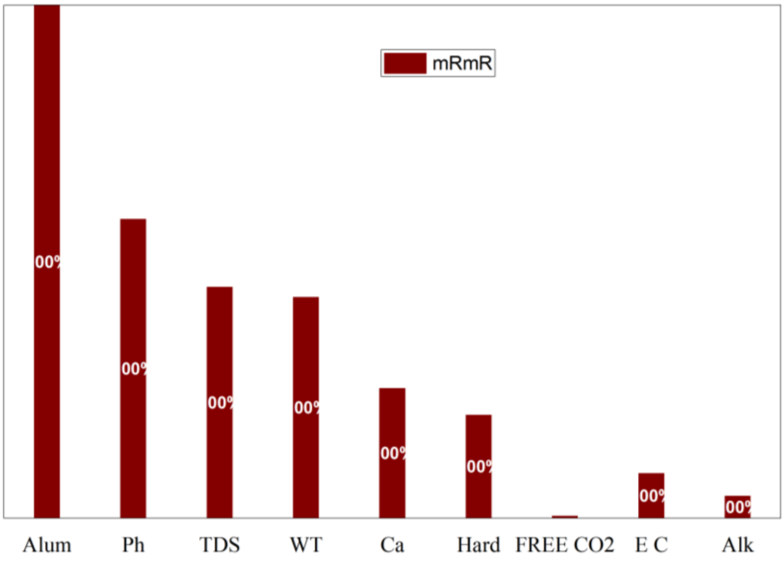


***Figure 6**: Illustrates the feature Ranking based on mRmR*

4.5 F-Test Feature Importance Analysis

The F-test ranking provides a statistically grounded evaluation of how strongly each physicochemical parameter contributes to explaining the variability in Turb at the TWTP. In this analysis, Alum concentration emerges as the dominant feature, exhibiting the highest F-score. This indicates a pronounced discriminatory power, reflecting the crucial role of coagulant dosing in shaping particle aggregation and subsequent Turb reduction. Such dominance is consistent with established coagulation–flocculation mechanisms where alum governs colloidal destabilization, directly influencing clarity outcomes [29]. Following Alum, Hard, Ca, and EC show progressively lower but still meaningful F-scores. Their positions in the ranking highlight their relevance in characterizing the ionic environment of the water column. Hard and Ca influence particle interactions and floc density, which in turn affects the settling dynamics and final Turb levels after treatment. Meanwhile, EC serves as an integrative indicator of dissolved ionic species, indirectly reflecting source-water characteristics that modulate coagulation requirements. The mid-tier variables free CO2, Alk, and pH demonstrate moderate statistical influence. Their F-scores reflect the buffering conditions and acid–base equilibria that govern coagulant effectiveness. Variations in Alk and pH can alter Alum hydrolysis species, changing floc formation behavior and thus contributing to Turb variability. Although their influence is not as strong as Alum or hardness, their contribution remains essential for process stability and optimal chemical performance. Towards the lower end of the ranking, WT and TDS register the smallest F-scores. While these factors affect fluid viscosity, microbial activity, and solubility dynamics, their direct statistical separation from Turb outcomes appears limited in the Tamburawa dataset. Their behavior suggests that short-term Turb fluctuations at the plant are less sensitive to temperature or dissolved solids than to coagulation chemistry and ionic balance. The Feature ranking based F-test was demonstrate in Figure 7.

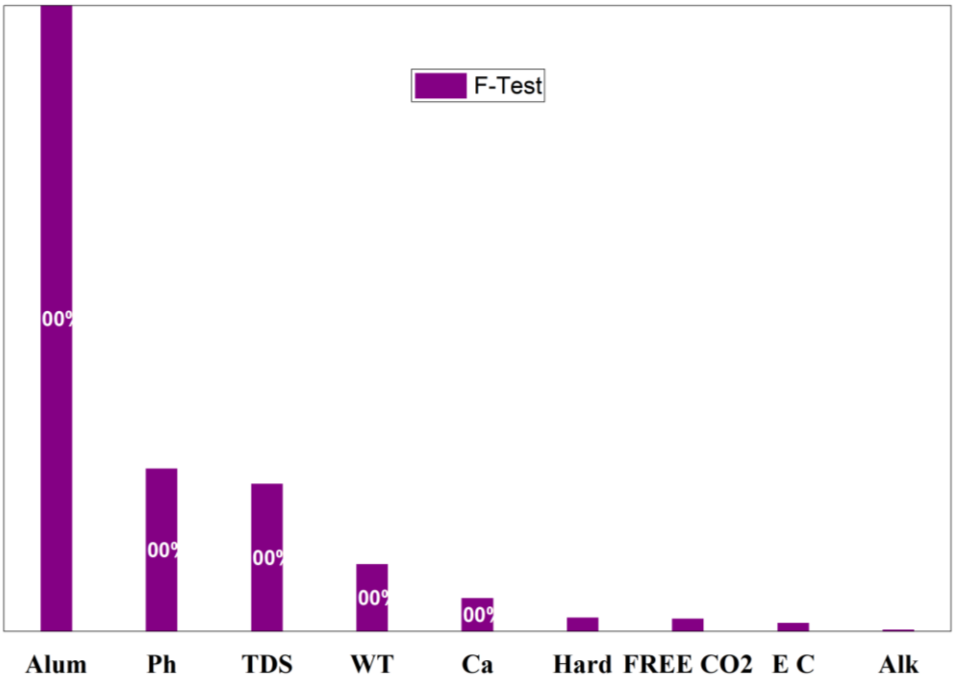


***Figure 7**: Illustrates the feature Ranking based F-Test*

4.6 SHAP Feature Influence Analysis

The SHAP evaluation provides a detailed, additive decomposition of how each predictor contributes to Turb fluctuations, revealing not only the ranking of influential variables but also the magnitude of their marginal impact on model output. In this analysis, E C stands unequivocally as the dominant driver of Turb predictions, with a mean absolute SHAP value of 0.02911, far exceeding that of any other parameter. This finding emphasizes the strong sensitivity of Turb dynamics to the ionic strength of raw water entering the TWTP. Higher EC typically reflects elevated concentrations of dissolved ions, which alter charge interactions surrounding suspended particles, change double-layer thickness, and influence the ease with which coagulants destabilize colloidal matter. The prominence of EC in the SHAP ranking is therefore consistent with mechanistic Turb behavior in semi-arid river systems, where upstream agricultural runoff, mineral dissolution, and seasonal hydrology play substantial roles. Following EC, Turb (lagged or raw measurement) appears as the second-most influential feature, with a SHAP value of 0.00856. This aligns with the well-known persistence of suspended particle loads in natural water bodies, where Turb often exhibits autocorrelation due to the slow settling velocity of fine sediments and organic colloids. The strength of this influence demonstrates that prior Turb conditions remain an important determinant of subsequent treatment outcomes, particularly under high-flow or disturbance periods. Free CO2, with a SHAP value of 0.00478, also exhibits a notable influence. Its position in the ranking reflects its role in shaping in-situ acidity and carbonate equilibria, thereby affecting both coagulant performance and particle stability. Elevated free CO2 generally shifts pH downward, modifies aluminum speciation when alum is applied, and affects the charge profile of suspended solids, all of which directly influence Turb removal efficiency. The SHAP magnitude indicates that fluctuations in free CO2 conditions have meaningful implications for day-to-day Turb dynamics in the plant, showing a SHAP value of 0.00221, which contributes moderately to Turb predictions through its control over fluid viscosity, settling behavior, and microbial interactions. Figure 8 indicates the feature ranking based on SHAP. Higher temperatures tend to enhance Brownian motion, influence coagulation kinetics, and accelerate biodegradation processes. Although its influence is lower than EC or free CO2, the SHAP ranking confirms that temperature variations still exert a non-negligible effect on Turb responses in the treatment system. The mid-lower tier of variables, Hardness, pH, Alum concentration, Calcium, and Alk, all register SHAP values below 0.002, indicating comparatively subtle contributions to prediction variability. Their reduced SHAP magnitudes do not imply irrelevance; rather, they suggest that, within the observed operating range of the Tamburawa plant, these parameters exhibit relatively stable dynamics or exert influence in conjunction with more dominant factors such as EC. For instance, alum dosing plays a central mechanistic role in Turb reduction, yet its SHAP value suggests that dosing levels were relatively consistent across the dataset, resulting in limited variability-driven impact on model predictions. Similarly, hardness and calcium influence coagulation conditions through ionic bridging and floc density, but their SHAP values show that day-to-day fluctuations were not strong enough to substantially drive model output on their own.

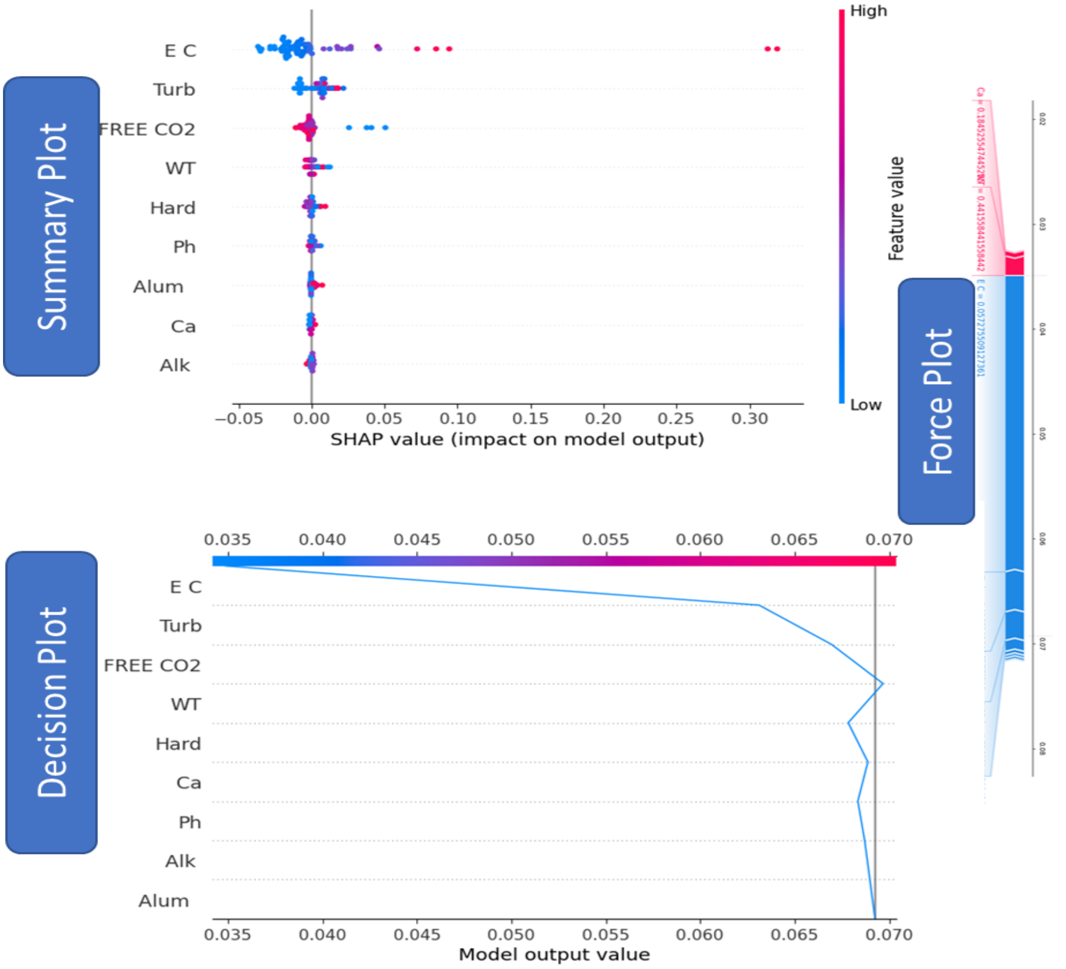


***Figure 8**: SHAP base feature Ranking*

4.7 Application of ML in Water Quality

ML has become an important tool for modelling WQ processes, particularly in situations where nonlinear interactions, multicollinearity, and complex feedbacks limit the usefulness of traditional regression-based approaches. In drinking-water treatment, ML models have been used to predict key quality indicators such as Turb, residual Ca, TDS, and composite indices, as well as to support optimization of coagulant dosing and process control [30]. Studies in surface- and groundwater systems have shown that algorithms such as ANN, SVM, DT, and GPR can capture nonlinear and interaction effects between physicochemical variables with higher accuracy than conventional linear models when trained on adequately pre-processed datasets. These approaches have been applied for forecasting river-water quality, groundwater salinity and nutrient loads, and for predicting treated-water parameters at plant outlets, demonstrating their suitability for real-time or near–real-time decision support in water utilities [31]. In the context of coagulation and Turb control, tree-based and NN models are particularly attractive because they can represent threshold behavior and interaction effects between variables such as EC, Hard, Ca, pH, and coagulant dose without requiring explicit specification of functional forms. Several studies have reported successful application of ANN, SVM, and BT for predicting Turb after coagulation, flocculation, and filtration [32]. Kernel-based methods such as SVM and probabilistic models such as GPR have the additional advantage of offering flexible, nonparametric function approximations with built-in regularization, which is useful when dealing with relatively small samples and correlated inputs typical of treatment-plant datasets [33].Furthermore, Figure 9 describe Violin Plot base on predicted and actual data set for both training and testing.

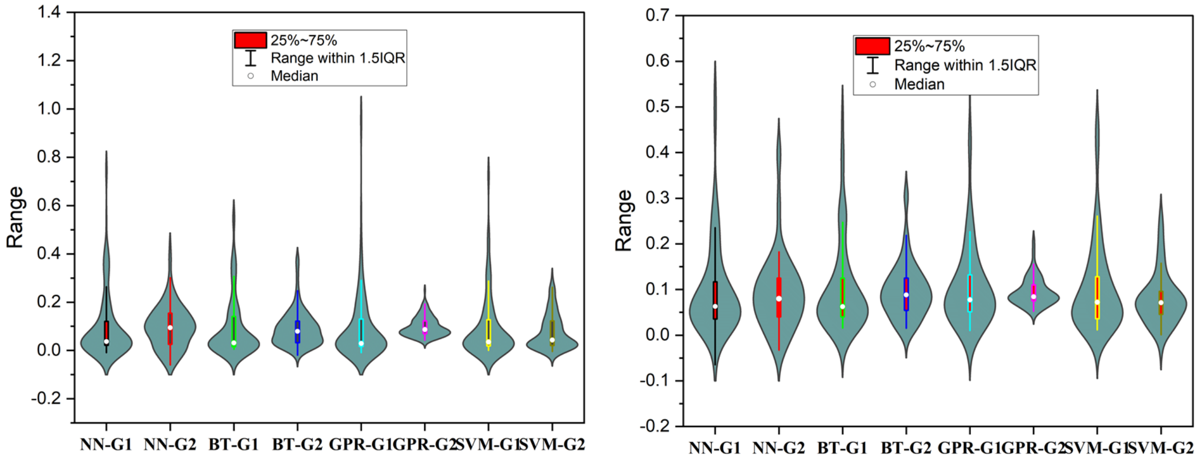


***Figure 9**: A Violin plot based on actual and predicted data instances*

4.8 Socioeconomic and Environmental Implications

The insights from this study have far-reaching implications for both environmental management and socioeconomic development. Turb directly influences the cost and complexity of water treatment; higher Turb levels demand increased filtration, chemical dosing, and energy consumption [12]. For communities reliant on groundwater for drinking and domestic use, poor Turb control translates into elevated financial burdens, infrastructure stress, and heightened public-health risks. This is particularly critical in regions where WT budgets are limited and dependence on groundwater is high. Environmentally, Turb affects aquatic life by reducing light penetration, impairing photosynthesis, and altering habitat quality. Persistent Turb may signal upstream erosion, land-use disturbances, or failing sanitation infrastructure, each carrying long-term ecological consequences. The identification of EC, free CO_2, and temperature as dominant drivers provides a strategic pathway for intervention, allowing authorities to focus monitoring efforts on parameters that offer the earliest warning signals of deteriorating WQ [34]. By combining statistical relevance with interpretable modeling, this study equips decision-makers with evidence-based insights needed for targeted regulation, resource allocation, and watershed protection strategies. These findings support sustainable groundwater management and align with broader global imperatives on environmental security and public health resilience.

5.0 Conclusion

This study demonstrates that Turb in groundwater systems can be reliably predicted using advanced data-driven modeling techniques supported by rigorous feature-attribution and statistical analyses. The boosted tree (BT-G1) model emerged as the most effective predictive framework, capturing the complex interactions between physicochemical variables and suspended-particle dynamics with the highest precision. The integration of correlation analysis, F-testing, and SHAP attribution produced a coherent interpretative structure. EC was consistently identified as the most critical predictor, followed by free CO_2 and temperature variables closely linked to both anthropogenic influence and natural geochemical processes. These findings underscore the multifactorial nature of Turb and highlight the importance of monitoring parameters that act as early indicators of watershed stress. Future research should expand the modeling framework by integrating hydrological variables such as flow rate, rainfall intensity, geological strata, and land-use patterns. Incorporating temporal sampling across seasons would also enhance model generalization and reveal how Turb responds to climatic variability. Additionally, incorporating sensor-network data could facilitate real-time prediction systems, enabling proactive water-treatment and watershed-protection strategies.

Competing Interests: The authors declare that they have no competing interests.

Data Availability Statement: The supported data associated with this researcher is available upon request from the corresponding author.

References

[1] S. I. Abba, G. Najashi, A. Rotimi, B. Musa, S. J. Kawu, and S. M. Lawan, "Jo ur na l P re of," *Results Eng.*, p. 100260, 2021, doi: 10.1016/j.rineng.2021.100260.

[2] A. Mosavi, M. Salimi, S. F. Ardabili, and T. Rabczuk, "State of the Art of Machine Learning Models in Energy Systems , a Systematic Review," 2019, doi: 10.3390/en12071301.

[3] A. Behzadi, E. Gholamian, P. Ahmadi, and A. Habibollahzade, "Energy , exergy and exergoeconomic (3E) analyses and multi-objective optimization of a solar and geothermal based integrated energy system," *Appl. Therm. Eng.*, vol. 143, no. April, pp. 1011–1022, 2018, doi: 10.1016/j.applthermaleng.2018.08.034.

[4] S. Baral, D. Kim, E. Yun, and K. C. Kim, "Energy, Exergy and Performance Analysis of Small-Scale Organic Rankine Cycle Systems for Electrical Power Generation Applicable in Rural Areas of Developing Countries," pp. 684–713, 2015, doi: 10.3390/en8020684.

[5] J. Usman *et al.*, "Enhanced desalination with polyamide thin-film membranes using ensemble ML chemometric methods and SHAP analysis," *RSC Adv.*, vol. 14, no. 43, pp. 31259–31273, 2024, doi: 10.1039/d4ra06078d.

[6] W. Zeng, Y. Qiu, Y. Huang, and Z. Luo, "Quantitative structure-retention relationship by databases of illegal additives," *J. Food Compos. Anal.*, vol. 122, no. June, p. 105500, 2023, doi: 10.1016/j.jfca.2023.105500.

[7] I. A. Mahmoud, U. J. Muhammad, S. J. Kawu, and M. M. Magaji, "Machine Learning-Based Wind Speed Estimation for Renewable Energy Optimization in Urban Environments : A Case Study in Kano State , Nigeria," no. March, 2024, doi: 10.52589/AJSTE-XKYBH2QI.

[8] F. Ligate *et al.*, "Geogenic contaminants and groundwater quality around Lake Victoria goldfields in northwestern Tanzania," *Chemosphere*, vol. 307, p. 135732, 2022, doi: https://doi.org/10.1016/j.chemosphere.2022.135732.

[9] X. Zhu *et al.*, "Effects of different types of anthropogenic disturbances and natural wetlands on water quality and microbial communities in a typical black-odor river," *Ecol. Indic.*, vol. 136, p. 108613, 2022, doi: https://doi.org/10.1016/j.ecolind.2022.108613.

[10] I. Mahmoud, "Deep learning LSTM and random forest ML-aided design tools for energy cooling capacity modeling." Apr. 07, 2025.

[11] O. Faith, A. Ali, and O. Oluwatosin, "An Assessment of Water Quality Status of Challawa River in Kano State, Nigeria 1," vol. 6, pp. 66–73, Jun. 2020.

[12] S. I. Abba *et al.*, "Optimization of Extreme Learning Machine with Metaheuristic Algorithms for Modelling Water Quality Parameters of Tamburawa Water Treatment Plant in Nigeria," *Water Resour. Manag.*, 2024, doi: 10.1007/s11269-024-04027-z.

[13] S. I. Abba *et al.*, "Journal of Water Process Engineering Emerging evolutionary algorithm integrated with kernel principal component analysis for modeling the performance of a water treatment plant," vol. 33, no. November 2019, 2020.

[14] I. A. Mahmoud, U. J. Muhammad, S. J. Kawu, and M. Mukhtar, "Enhancing Energy Demand Prediction Using Elman Neural Network and Case Study Enhancing Energy Demand Prediction Using Elman Neural Network and Support Vector Machine Model : A Case Study in Lagos State , Nigeria," no. November, 2024, doi: 10.37256/aie.5220244396.

[15] I. A. Mahmoud, A. A. Wakili, A. M. Danjuma, and A. Y. Sada, "Deep learning LSTM and random forest ML-aided design tools for energy cooling capacity modeling," no. March, pp. 1–13, 2025.

[16] U. J. Muhammad, I. I. Aminu, I. A. Mahmoud, U. U. Aliyu, and A. G. Usman, "AI in Civil Engineering An improved prediction of high - performance concrete compressive strength using ensemble models and neural networks," *AI Civ. Eng.*, 2024, doi: 10.1007/s43503-024-00040-8.

[17] I. A. Mahmoud, H. M. Umar, M. Sa, and M. G. Abdullahi, "Water-Energy Nexus for Global Sustainability : A Comprehensive Review , Challenges , Innovations , and Strategic Solutions," vol. 1, no. August, pp. 26–40, 2025.

[18] M. S. Gaya, S. I. Abba, A. M. Abdu, and A. I. Tukur, "Estimation of water quality index using artificial intelligence approaches and multi-linear regression," vol. 9, no. 1, pp. 126–134, 2020, doi: 10.11591/ijai.v9.i1.pp126-134.

[19] S. I. Abba *et al.*, "Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination," *J. Hydrol.*, vol. 587, p. 124974, 2020, doi: https://doi.org/10.1016/j.jhydrol.2020.124974.

[20] W. Zieliński, E. Korzeniewska, M. Harnisz, J. Drzymała, E. Felis, and S. Bajkacz, "Wastewater treatment plants as a reservoir of integrase and antibiotic resistance genes – An epidemiological threat to workers and environment," *Environ. Int.*, vol. 156, p. 106641, 2021, doi: https://doi.org/10.1016/j.envint.2021.106641.

[21] E. Jumin, F. B. Basaruddin, Y. B. Yusoff, S. D. Latif, and A. N. Ahmed, "Solar radiation prediction using boosted decision tree regression model : A case study in Malaysia Solar radiation prediction using boosted decision tree regression model : A case study in Malaysia," no. October 2023, 2021, doi: 10.1007/s11356-021-12435-6.

[22] U. U. Aliyu *et al.*, "Biomass and Bioenergy Optimizing biomedical waste generation modeling using quantum machine learning and economic development indicators," vol. 204, no. July 2025, 2025.

[23] I. A. Mahmoud *et al.*, "Multi-model environmental modelling of energy-exergy efficiency using GUI-based aided design tools integrated with dependency feature analysis," *Hybrid Adv.*, vol. 10, p. 100493, 2025, doi: https://doi.org/10.1016/j.hybadv.2025.100493.

[24] Z. Liang, Y. Li, Y. Hu, B. Li, and J. Wang, "A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework," *Theor. Appl. Climatol.*, vol. 133, no. 1–2, pp. 137–149, 2018, doi: 10.1007/s00704-017-2186-6.

[25] O. Sureiman and C. Mangera, "F-test of overall significance in regression analysis simplified," *J. Pract. Cardiovasc. Sci.*, vol. 6, p. 116, Jan. 2020, doi: 10.4103/jpcs.jpcs_18_20.

[26] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, pp. 189–203, 2018, doi: https://doi.org/10.1016/j.jbi.2018.07.014.

[27] A. M. Jibrin *et al.*, "Influence of membrane characteristics and operational parameters on predictive control of permeance and rejection rate using explainable artificial intelligence (XAI) Analysis of Variance," vol. 2, no. July 2024, 2025, doi: 10.1016/j.nexres.2024.100100.

[28] S. M. Tabatabaei, M. Asadian-Pakfar, and B. Sedaee, "Well placement optimization with a novel swarm intelligence optimization algorithm: Sparrow Search Algorithm," *Geoenergy Sci. Eng.*, vol. 231, p. 212291, 2023, doi: https://doi.org/10.1016/j.geoen.2023.212291.

[29] W. R. Zwick and W. F. Velicer, "Comparison of Five Rules for Determining the Number of Components to Retain," *Psychol. Bull.*, vol. 99, no. 3, pp. 432–442, 1986, doi: 10.1037/0033-2909.99.3.432.

[30] E. K. Nti *et al.*, "Water pollution control and revitalization using advanced technologies: Uncovering artificial intelligence options towards environmental health protection, sustainability and water security," *Heliyon*, vol. 9, no. 7, p. e18170, 2023, doi:

https://doi.org/10.1016/j.heliyon.2023.e18170.

[31] T. A. Adesakin *et al.*, "Assessment of bacteriological quality and physico-chemical parameters of domestic water sources in Samaru community, Zaria, Northwest Nigeria," *Heliyon*, vol. 6, no. 8, p. e04773, 2020, doi: https://doi.org/10.1016/j.heliyon.2020.e04773.

[32] A. M. Jibrin *et al.*, "Machine learning predictive insight of water pollution and groundwater quality in the Eastern Province of Saudi Arabia," *Sci. Rep.*, pp. 1–16, 2024, doi: 10.1038/s41598-024-70610-4.

[33] S. I. Abba, S. Jasim, and J. Abdullahi, "ScienceDirect ScienceDirect River water modelling prediction using multi-linear regression , artificial neural network , and adaptive neuro-fuzzy inference system techniques," *Procedia Comput. Sci.*, vol. 120, pp. 75–82, 2018, doi: 10.1016/j.procs.2017.11.212.

[34] Y. Sun, D. Wang, L. Li, R. Ning, S. Yu, and N. Gao, "Application of remote sensing technology in water quality monitoring: From traditional approaches to artificial intelligence," *Water Res.*, vol. 267, p. 122546, 2024, doi: https://doi.org/10.1016/j.watres.2024.122546.