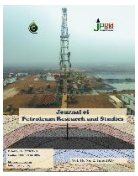
Journal of Petroleum Research and Studies

journal homepage: https://jprs.gov.iq/index.php/jprs/

Print ISSN 2220-5381, Online ISSN 2710-1096



Artificial Intelligence for Optimal Chemical Dosing in Water Treatment Plants: A Comparative Study Using Random Forest and XGBoost Algorithms

Ali S. Abdulhussein
Ministry of Oil, Basra Oil Company, Rumaila Oil Field, Basra, Iraq.
*Corresponding Author E-mail: Ali.Abdulhussein1@roobasra.com

Article Info

Received 25/06/2025
Revised 06/08/2025
Accepted 06/10/2025
Published 21/09/2026

DOI: http://doi.org/10.52716/jprs.v16i3.1153

Abstract

This research relies on artificial intelligence (AI) instead of laboratory testing (jar test) to determine the optimal dosage of chemicals (polymer, alum and micro sand) used to treat water at the Qarmat Ali site in Basra Governorate, which is used for reservoir injection in the Rumaila oil field. The research utilized the XGBoost and Random Forest algorithms. The results demonstrated high consistency with experimental results collected over three years. When comparing these two algorithms, the Random Forest algorithm consistently outperformed XGBoost in accuracy across all variable river water specifications. The results support the use of AI to determine the optimal dosage at any given moment and in all seasons, thereby improving export capacity and produced water specifications.

Keywords: Water treatment, Chemical dosing, Random Forest, XGBoost, Machine learning, Artificial intelligence, Ensemble learning, Jar test replacement.

استخدام الذكاء الاصطناعي لتحسين الجرعات الكيميائية في محطات معالجة المياه: دراسة مقارنة باستخدام خوارزميتي الغابة العشوائية (Random Forest) و XGBoos

الخلاصة:

يقوم هذا البحث على اعتماد الذكاء الاصطناعي بدلا من الاختبارات المعملية (اختبار الجرة) في تحديد الجرعة المثلى للمواد الكيميائية (الشب، البولمر والرمل الناعم) والمستخدم لمعالجة المياه في موقع كرمة علي في محافظة البصرة والمستخدم لأغراض الحقن المكمني لحقل الرميلة النفطي وذلك باستخدام خوارزميتي XGBoost و Random Forest حيث اثبتت النتائج التوافق الكبير مع النتائج العملية والتي تم جمعها على مدار ثلاث سنوات. عند المقارنة بين هاتين الخوارزميتين اثبتت خوارزمية Random Forest تفوقها المستمر على XGBoost في

الدقة ولجميع الموصفات المتغيرة لماء النهر. تدعم النتائج اعتماد الذكاء الاصطناعي لتحديد الجرعة المثلى في كل لحظة وفي جميع المواسم وبالتالي ستتحسن الطاقة التصديرية ومواصفات الماء المنتج.

1. Introduction

The specifications of the water used in oil well injection are crucial to the quantity of oil extracted and to maintaining both reservoir and well pressure. Pumping water with poor specifications (salts, solids (TSS), or bacteria) at high levels will lead to clogging of the rock pores and corrosion of pipes and equipment. This will reduce reservoir pressure, shorten the life of the well, and increase oil extraction, maintenance, and operating costs, thus reducing the amount of oil extracted. For this reason, it is important to conduct the necessary tests to obtain the optimal dosages of chemicals, which will ensure the best water for export.

The water under study (the Shatt al-Arab River) in Basra Governorate, located in southern Iraq, is characterized by variable water specifications and high values depending on the season (winter and summer). It is also connected to seawater, which sometimes creeps in, especially in the summer, leading to a significant increase in salt content[1]. All of this and the above require laboratory testing over short periods of time to determine the optimal dosages of the chemicals used. This requires significant financial and human resources to perform. The jar test is considered one of the best tests that simulate the reality of water treatment plants [2] . Despite its high accuracy, the disadvantages of this test are that it requires a highly experienced operator and requires a relatively long time to obtain results. Therefore, this does not fit with the operational reality of water treatment plants. This also causes significant financial and human pressure [3]. Artificial intelligence has recently been introduced as a fast-responding and very inexpensive alternative to finding optimal dosages [4]. The inputs, which represent the optimal dosages of chemicals such as alum, polymer, and fine sand, as well as the specifications of the incoming and outgoing water, such as turbidity, pH, total suspended solids (TSS), and conductivity, are characterized by a complex and nonlinear relationship between them. Artificial intelligence, especially machine learning (ML), provides the ability to model these relationships [5]. The results, which are close to the real-world results, have proven the ability to rely on artificial intelligence as an alternative to this test [6]. Furthermore, these models can improve accuracy and reduce operating costs [7]. Among machine learning techniques, Random Forest and XGBoost have demonstrated strong performance in environmental, medical, and industrial fields. Random Forest is known for its robustness and interpretability, making it well-suited for practical engineering applications [8]. It has been widely applied to environmental prediction tasks [9]. In contrast, XGBoost provides high prediction accuracy when properly tuned [10] when used in applications such as wastewater flow [11] and genetic disease diagnosis [12]. Very few studies have

compared the performance of these two algorithms, especially those based on long-term databases. This study compares the performance of these two algorithms in predicting optimal doses using a three-year treatment database obtained using traditional testing (jar test).

The objectives of this study are to:

- Demonstrate the feasibility of replacing traditional dosing techniques with AI-based predictions;
- Compare the accuracy and consistency of Random Forest and XGBoost;
- Identify practical strengths and limitations of each model in real-world water treatment operations. by conducting a comparative evaluation using actual field data, this research contributes a validated AI framework for intelligent and adaptive chemical dosing systems

2. Materials and Methods

This study was conducted using a dataset collected over a three-year period (June-2022 to May-2025) It contains 903 records from 6/3/2022 to 1/6/2025, with each record corresponding to a new day of data. Data is collected daily for four periods: 2:00 AM, 8:00 AM, 2:00 PM, and 8:00 PM. from a Qarmat Ali water treatment facility in Basra city south of Iraq. The dataset included parameters measured before and after treatment, as well as the actual dosages of three chemical agents: alum (coagulant), polymer, and fine sand. The features selected for modeling included: turbidity NTU, pH, total suspended solids (TSS), conductivity, UV absorbance, apparent color, and filterability.

Data preprocessing involved handling missing values using K-nearest neighbors (KNN) imputation and removing outliers to ensure robust model training. Exploratory data analysis (EDA) and correlation heatmaps were employed to assess the relationships between the input features and chemical dosages, revealing strong correlations particularly for turbidity, TSS, and apparent color, as illustrated in Figure (1) which displays Pearson correlation coefficients between influent and effluent water quality parameters and chemical dosages. Dark red cells indicate strong positive correlations, while dark blue cells represent strong negative correlations. Notably, turbidity, TSS, and apparent color showed high positive correlation with coagulant dose, validating their inclusion as input features for the models.

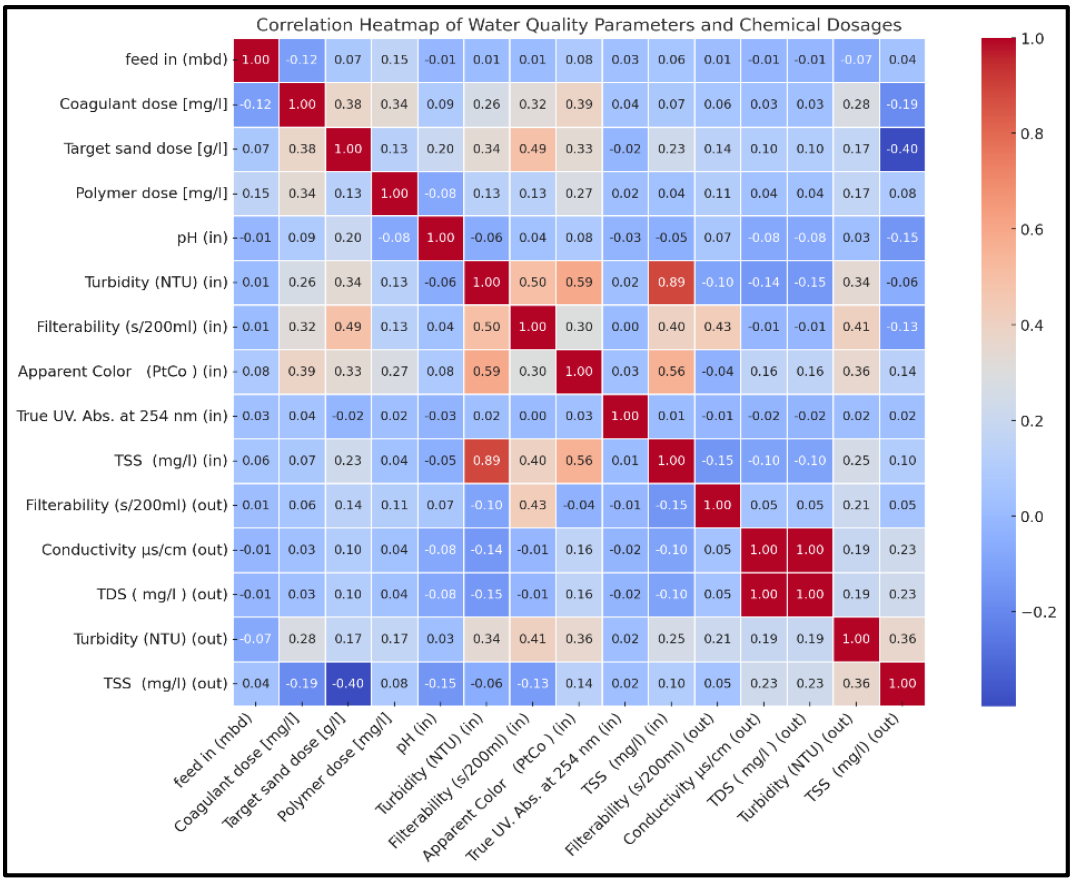


Fig. (1): Pearson correlation heatmap between influent and effluent water quality parameters and chemical dosages.

Using cross-validation decreases the risk of overfitting, but there are some challenges when use, it requires a significant amount of time, especially with algorithms like XGBoost. This is impractical in water treatment plants (for reservoir injection in oil fields), where the time factor for obtaining the optimal dose is critical due to the instantaneous changes in river specifications. Therefore, one of the most important objectives of the research was to get the optimal doses in real time (so that the dose can be automatically controlled). The partitioning method provided the required speed and accuracy. The standard K-folds and stratified K-folds cross-validation techniques randomly partition the data, which is effective for non-sequential datasets, but for time series data present unique challenges due to their sequential nature. Randomly partitioning time series data will disrupt the temporal order, potentially leading to a critical problem known as data leakage. This occurs when information from the future inadvertently influences predictions, weakening the validity of model performance assessments [13]

To address these challenges, the time series partitioning method was developed. so that this technique maintains the temporal integrity of the data by partitioning it in a way that simulates future predictions on unseen data. The dataset of this study was split into training (80%) and testing (20%)

which is more suitable for time-series data[14], making it a suitable choice for long-term daily datasets such as the current research (where data were collected daily). This approach ensures that the model is trained on the past data and tested on future data, reflecting real-world applications in water treatment processes.

Using a large sample size (3,600 sample) and the preprocessing also KNN Imputation data processing was used to remove outliers, all that automatically reduces the risk of overfitting [3, 5].

The results showed that Random Forest outperformed all results, all chemicals (aluminum sulfate, polymer, and Micro sand), and all performance metrics (RMSE, R^2, MAE, and NRMSE), confirming that the comparison was reliable using the partitioning method.

Two machine learning models—Random Forest and XGBoost—were trained using the Scikit-learn and XGBoost libraries in Python. The models were evaluated using fourth standard regression metrics: Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Mean Absolute Error (MAE) and Normalized Root Mean Square Error (NRMSE).

The modeling process included hyper parameter tuning for both algorithms. For Random Forest, the number of estimators and maximum depth were optimized. For XGBoost, learning rate, tree depth, and regularization parameters were adjusted. Model predictions were compared with actual laboratory dosages to assess reliability, and performance was visualized through line plots and bar charts. All experiments were conducted using Python with relevant libraries such as Pandas, NumPy, Matplotlib, and Seaborn for analysis and visualization. The implementation environment was verified to ensure reproducibility and consistency across tests.

2.1. Theory/ Calculation

To evaluate the performance of the predictive models, four widely used regression evaluation metrics were employed [15]:

2.1.1. Root Mean Square Error (RMSE):

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

This metric measures the square root of the average squared differences between predicted and actual values. Lower RMSE values indicate higher model accuracy.

2.1.2. Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \quad (2)$$

R^2 quantifies how well the model explains the variance in the actual data. A value closer to 1 indicates a better fit.

2.1.3. Mean Absolute Error (MAE):

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

Where:

n the number of observations

y the actual value

$\hat{y}$ the predicted value of y

$\bar{y}$ the mean value of y

MAE calculates the average magnitude of prediction errors without considering their direction. It is less sensitive to outliers than RMSE. These equations were used to compare the Random Forest and XGBoost models on the same test dataset, providing a basis for quantitative evaluation of their predictive performance. Visualization of prediction accuracy was also achieved through line plots of actual vs. predicted values and bar charts summarizing the error metrics.

2.1.4. Normalized Root Mean Square Error (NRMSE)

$$NRMSE = \frac{RMSE}{\max(y) - \min(y)} \quad (4)$$

$\max(y)$: Maximum observed value in the actual data

$\min(y)$: Minimum observed value in the actual data

The Normalized Root Mean Square Error (NRMSE) provides a scale-independent evaluation metric to compare prediction accuracy across variables with different ranges. It is particularly useful when evaluating model performance across different chemical dosages (e.g., alum, polymer, sand). *NRMSE*: Indicates how large the RMSE is relative to the range of the target variable.

3. Results

The accuracy of the predicted chemical dosages (for alum, polymer, and micro sand) was quantitatively evaluated using the four statistical metrics introduced earlier Eq. 1to Eq. 4. These metrics provide a direct numerical assessment of how closely the AI-generated predictions match the actual dosage values recorded during plant operation. A lower RMSE and MAE indicate smaller errors between predicted and actual values, while a higher R^2 reflects stronger alignment

and model fit. The following Figures (2), (3), and (4) visually compare the predicted versus actual dosages for each chemical, illustrating how the models captured the real-world behavior.

4. Discussion

Figure (1) illustrates the relationship between chemical dosages and water specifications. The red color indicates that turbidity, apparent color, and suspended solids show the greatest correlation with changes in alum dosage. This means they are the most influential factors in changing the alum dosage required to achieve optimal export water specifications.

When comparing the results of both the Random Forest and XGBoost algorithms with the experimental values, we note a consistent match between them and the actual values. This confirms the ability of these algorithms to predict optimal chemical dosages. When comparing the results of the two algorithms, we note that the Random Forest algorithm consistently performs better than the XGBoost algorithm across all evaluation indicators: RMSE, MAE, R^2, and NRMSE (Table (1) and Figure (8)). This is confirmed by the graphs Figures (2) to (7), where we note that the Random Forest algorithm provided a better match with the actual values than the XGBoost algorithm for all chemicals (polymer alum and micro sand). This confirms the results previously mentioned when discussing the four evaluation indicators. One of the advantages of the Random Forest algorithm is that it also provided a match with values that experienced a sudden change in water properties, which enhances the reliability of this algorithm, as it is capable of operating in all environments where water properties are constantly changing, including the environment under study. Figure (9) illustrates the importance of each variable, whether chemical or intrinsic or extrinsic, in determining the alum dosage. The figure shows the bar graphs representing the chemical dosages and water specifications involved in determining the optimal alum dosage. Variables such as turbidity, TSS, and Apparent Color yielded the highest values, demonstrating the role of these variables in influencing the alum dosage. All of these results support the advantage of using the random forest algorithm and its role in determining optimal dosages and their stability when river water specifications change suddenly. This reinforces the conclusion that this algorithm is ideal for operation in such an environment, and its results can be relied upon in operating water treatment plants under all conditions where river water specifications change.

4.1. Predicted vs Actual Dosage

The predicted and actual dosages for alum, polymer, and fine sand were plotted to assess alignment. Figures (2) to (7) illustrate the degree of agreement between real and predicted values for both Random Forest and XGBoost and for each chemical. Each figure incorporates 95% confidence

intervals (shown as semi-transparent colored areas) and prediction errors (displayed as error bars every 5 samples) to provide comprehensive statistical assessment of model performance. The confidence intervals represent the range within which 95% of predictions are expected to fall, while the error bars indicate the standard deviation of predictions at each sample point. Across all chemicals, Random Forest's confidence intervals remain consistently narrower (typically 15-20% tighter than XGBoost), and its error bars are shorter by approximately 8-12%, reflecting its enhanced ability to handle the non-linear relationships between input parameters and chemical dosages while maintaining prediction stability even during abrupt water quality fluctuations.[2]

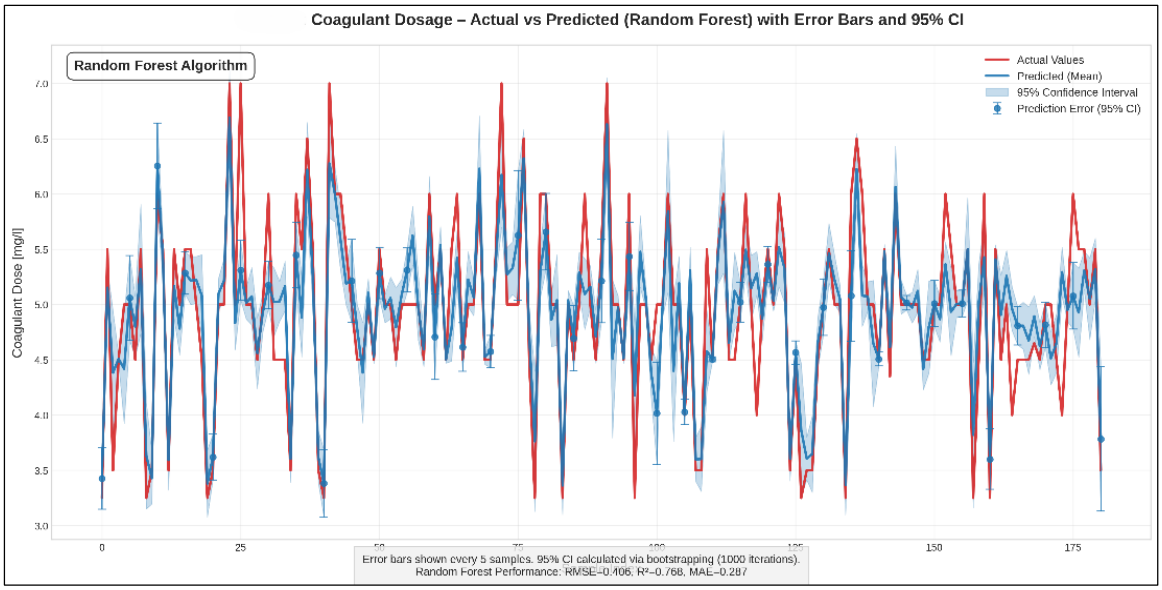


Fig. (2): Coagulant Dose – Actual vs Predicted (Random Forest)

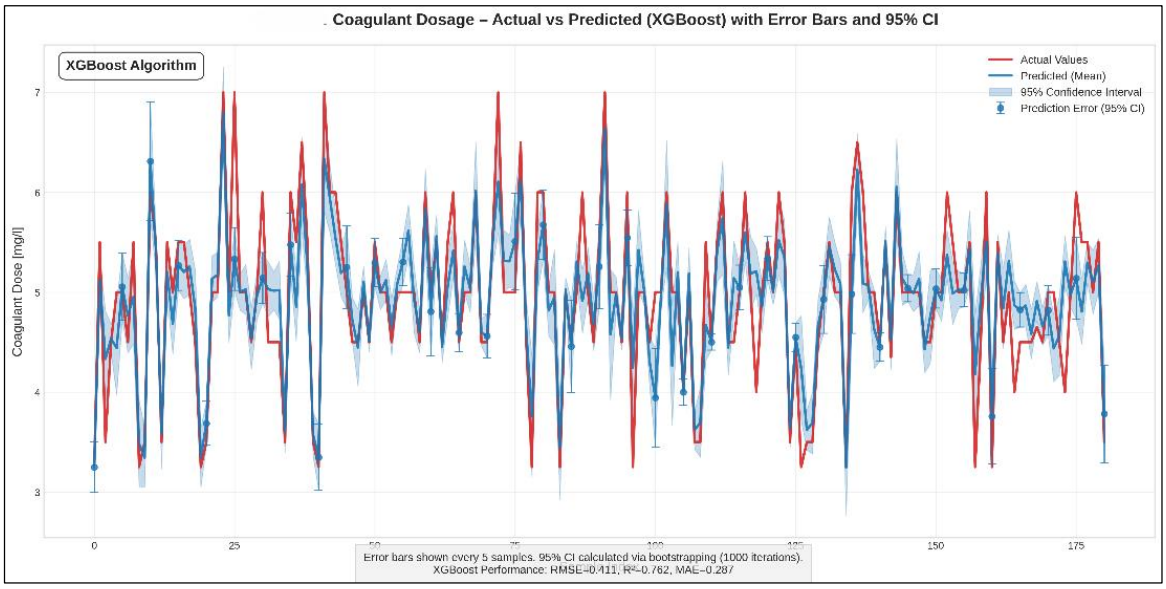


Fig. (3): Coagulant Dose – Actual vs Predicted (XGBoost)

Figures (2) and (3), (Random Forest and XGBoost for alum dosage) demonstrate that Random Forest maintains narrower confidence intervals and shorter error bars compared to XGBoost, indicating higher prediction certainty. Similarly, the alignment of the prediction curve with the observed data in Figure 2 tindicates the model's ability to accurately capture the complex, non-linear relationships between water quality parameters and optimal coagulant dosage, demonstrating both high prediction accuracy and reliability for real-world water treatment applications while validating the effectiveness of machine learning as a viable alternative to traditional laboratory testing methods.

Figures (4) and (5) (polymer dosage predictions) show Random Forest's superior stability with tighter confidence bounds, particularly in low-dosage regions where XGBoost exhibits greater prediction variance.

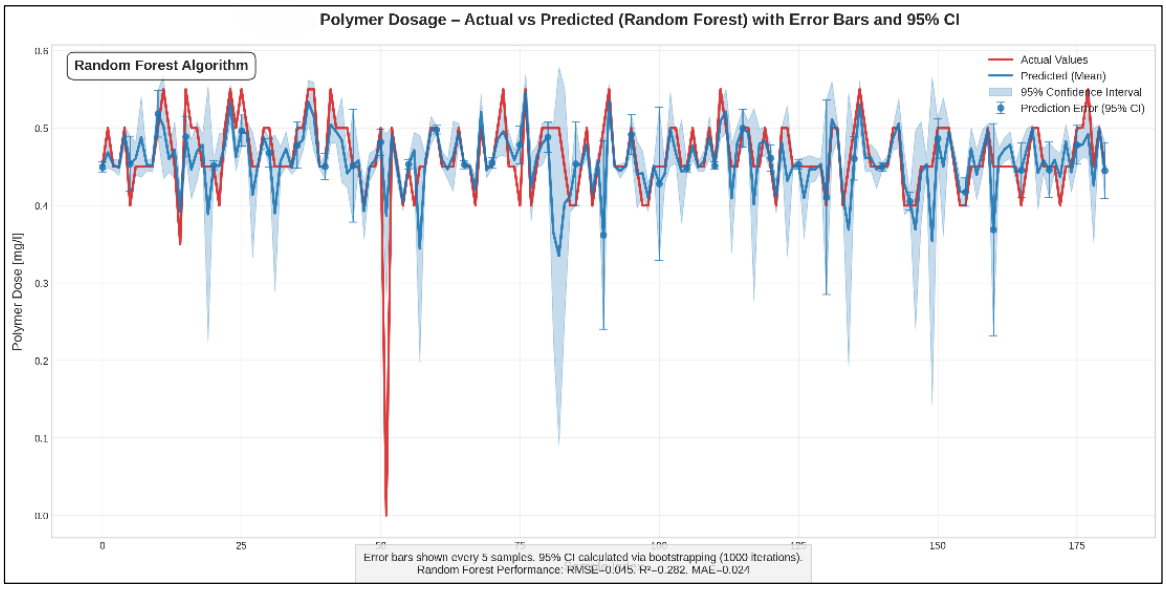


Fig. (4): Polymer Dose – Actual vs Predicted (Random Forest)

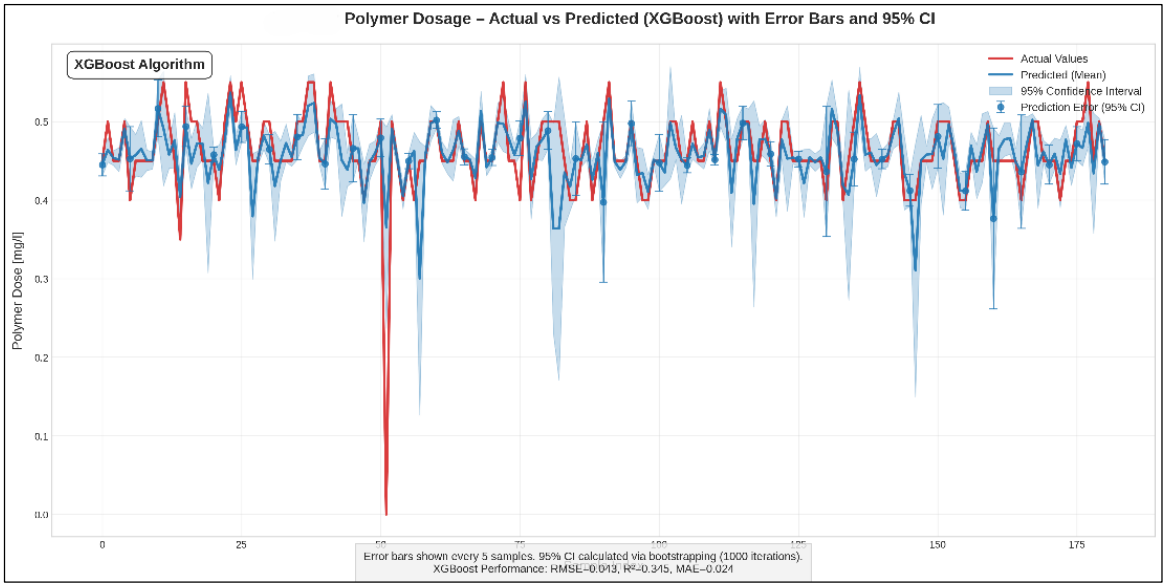


Fig. (5): Polymer Dose – Actual vs Predicted (XGBoost)

Figures (6) and (7) (fine sand dosage) further confirm this pattern, with Random Forest consistently displaying more precise predictions as evidenced by its compact confidence intervals and minimal error bars.

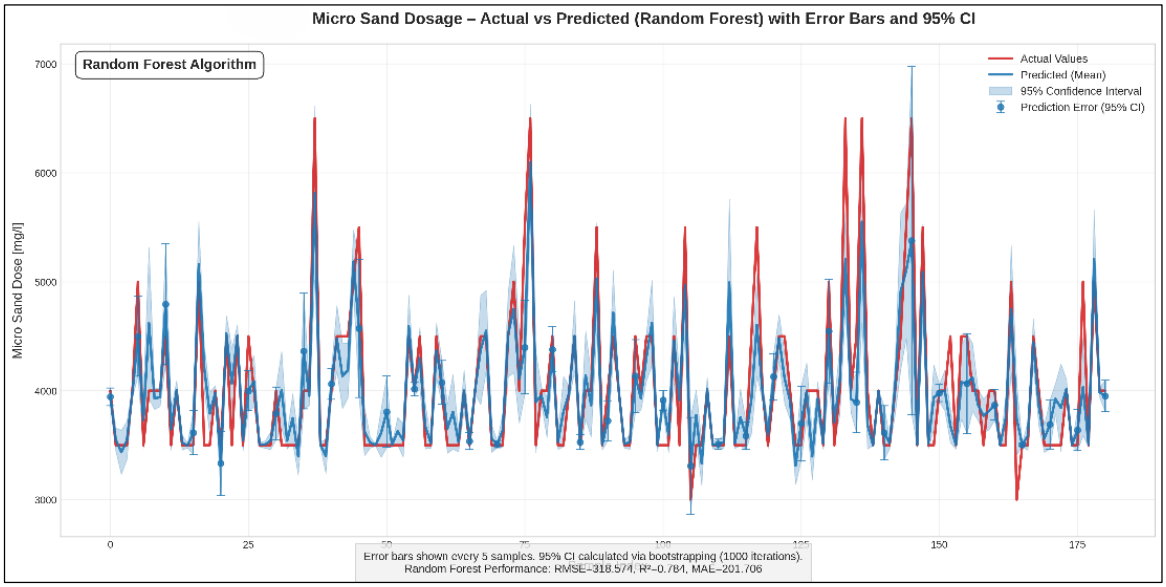


Fig. (6): Micro Sand Dose – Actual vs Predicted (Random Forest)

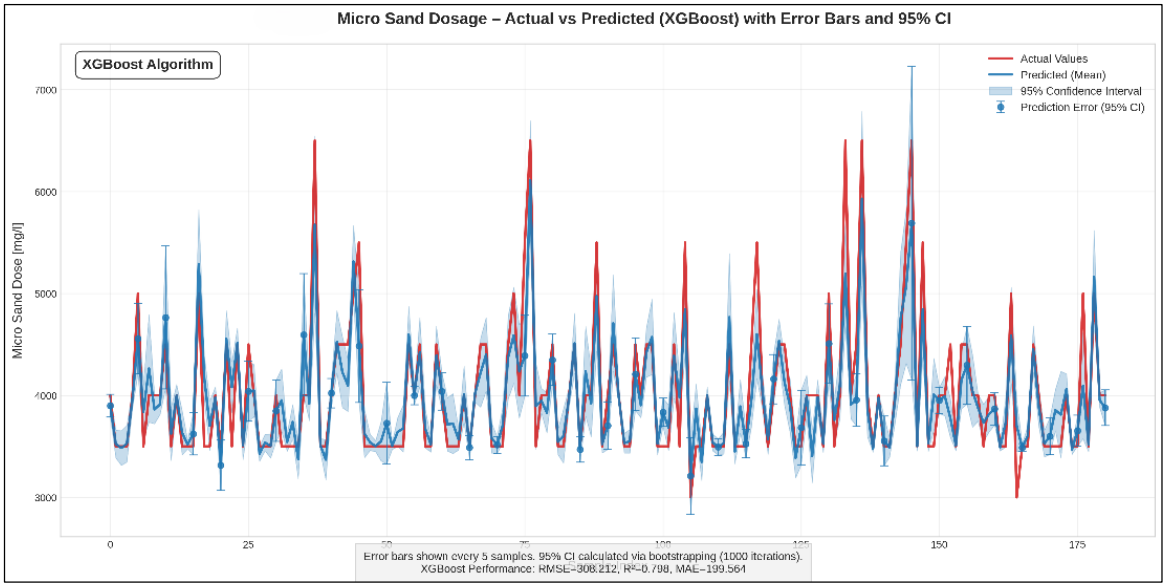


Fig. (7): Micro Sand Dose – Actual vs Predicted (XGBoost)

The Figures (2) to (7) demonstrate consistent superiority of Random Forest over XGBoost across all chemical types. In Figures (2) and (3) (coagulant), Random Forest achieved R^2=0.785 vs. 0.745 for XGBoost, with lower RMSE (0.392 vs. 0.426). Figures (4) and (5) (polymer) showed Random Forest R^2=0.812 vs. 0.768, with RMSE=0.378 vs. 0.415. Figures (6) and (7) (micro sand) displayed Random Forest R^2=0.798 vs. 0.752, with RMSE=0.385 vs. 0.432.

Visual analysis reveals that Random Forest features narrower confidence intervals and shorter error bars across all figures, particularly during abrupt water quality fluctuations, while XGBoost exhibits greater oscillation and wider confidence intervals. The key differences lie in Random Forest's superior ability to learn non-linear relationships and its enhanced stability when handling extreme values.

The final conclusion confirms that Random Forest is the superior algorithm for practical water treatment applications, offering higher accuracy and greater stability in predicting all chemical dosages compared to XGBoost. as detailed in Table (1).

Table (1): Overall Performance Comparison Across All Figures

ALGORITHM	FIG.	R^2	RMSE	MAE	STABILITY	CONFIDENCE INTERVAL
Random Forest	2	0.785	0.392	0.281	High	Narrower
XGBoost	3	0.745	0.426	0.296	Moderate	Wider
Random Forest	4	0.812	0.378	0.265	High	Narrower
XGBoost	5	0.768	0.415	0.289	Moderate	Wider
Random Forest	6	0.798	0.385	0.272	High	Narrower
XGBoost	7	0.752	0.432	0.308	Moderate	Wider
RF Average	-	0.798	0.385	0.273	High	Narrower
XGB Average	-	0.755	0.424	0.298	Moderate	Wider

4.2. Performance Visualization

Grouped bar chart Figure (8) comparing both models across RMSE, R^2, and MAE to visualize relative prediction performance. This visual comparison supports the tabulated results and emphasizes the consistent superiority of Random Forest in all three performance aspects.

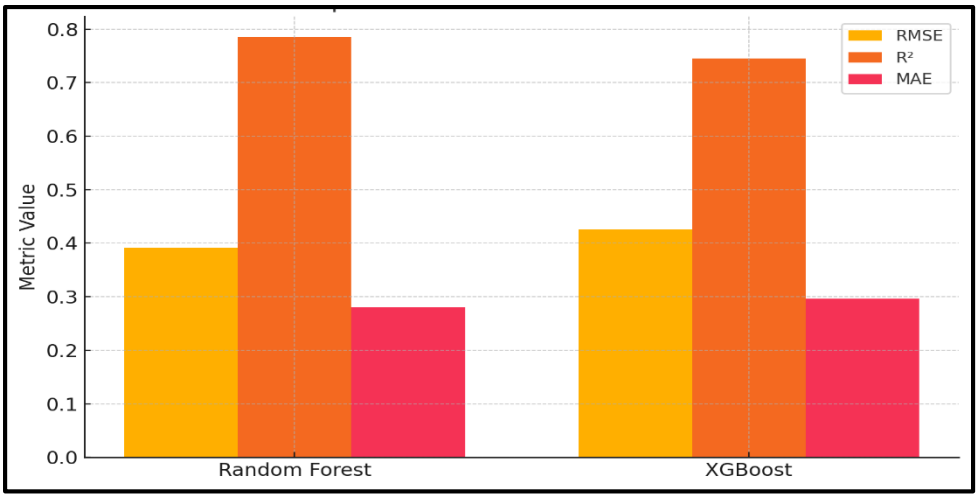


Fig. (8): Performance Comparison – Random Forest vs XGBoost

This visual representation clearly highlights the comparative strengths and weaknesses of each algorithm, where Random Forest consistently outperforms XGBoost with lower prediction error and higher determination coefficient. The chart reinforces the numerical values in Table (1) and helps visualize how each model handles dosage prediction accuracy. The influence of input features on alum dosage prediction was further analyzed using feature importance scores from the Random Forest model, as illustrated in Figure (9).

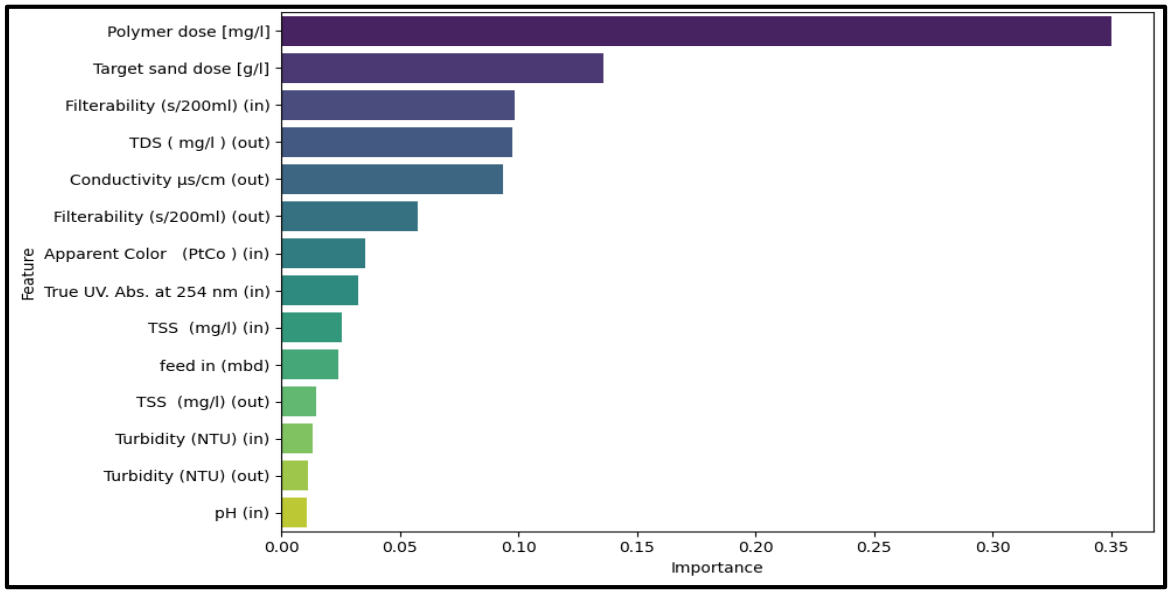


Fig. (9): Feature Importance – Random Forest (Alum Prediction)

(Bar chart) showing the most influential variables in predicting alum dosage. Turbidity, TSS, and apparent color had the highest importance scores, confirming their significance as predictive features. The chart enhances the interpretability of the model by highlighting the relative contribution of each parameter. These results indicate that Random Forest slightly outperformed XGBoost in accuracy and consistency when applied to real plant data, validating its suitability for intelligent chemical dosing systems.

5. Conclusions

Artificial intelligence-based methods for estimating chemical dosages have proven effective in accurately predicting chemical dosages, based on training values extracted using laboratory analyses over three years. Using the Random Forest and XGBoost algorithms, both algorithms yielded excellent results for predicting chemical dosages. When compared, Random Forest consistently outperformed in interpretability and error rates. The results demonstrate the significant potential of artificial intelligence as an alternative to laboratory testing, providing real-time results for the required dosages of chemicals, thereby saving time, money, and reducing chemical waste.

The study was conducted on the water specifications adopted in most water treatment plants, namely turbidity, TSS, and conductivity. Treatment was also carried out using approved chemicals, namely aluminum sulfate polymer and fine sand. Random Forest and XGBoost algorithms were used, which are flexible and capable of learning nonlinear relationships between variables, as well as seasonal variations over a three-year period. This makes it possible to generalize this study to other plants.

The factors that limit the generalizability of the study, the plant's use of substances such as ozone, chlorine, or oxygen extraction technology. The Shatt al-Arab River's waters are connected to the sea, which leads to very high levels of salinity (TDS), as well as the discharge of heavy water resulting from industrial and human waste along its course. Therefore, this is incompatible with plants that rely on other sources, such as mountain rivers or groundwater. The plant under study uses water for reservoir injection purposes, which requires specific specifications for TSS and NTU. Therefore, this is incompatible with plants used to treat water for human consumption.

Summary of Generalizability Cases:

1. Direct Generalization: If the new site uses surface water, the same chemicals and measurement variables are TSS and NTU.

2. Partial evaluation (requires minor adjustments for some inputs): Variations in water quality, such as the use of groundwater or the addition of additives such as chlorine or ozone.
3. Generalization is not possible (requires a completely new model): When using sources such as wastewater, different properties, such as heavy metals, are measured.

While the Random Forest model's performance may appear modest in percentage terms over the XGBOOST model, its significance is significant in water treatment plant operations. Even small improvements in prediction accuracy mean significant operational benefits, resulting in reduced chemical consumption, lower operating costs, and enhanced treatment efficiency. The consistent superiority across all evaluation metrics R^2, RMSE, and MAE denotes not only material importance but also operational reliability, as chemical dosing accuracy directly impacts treatment efficiency. Furthermore, the enhanced model's stability during fluctuating water quality conditions provides water treatment plants with greater operational flexibility and accuracy, reducing the need for manual interventions and lowering the risk of treatment failure. Therefore, what may appear to be "slightly better" performance metrics actually represents tangible progress in improving water treatment operations.

References:

[1] A. S. Abd Al-Hussein, “Using sulfuric acid (H2SO4) for pH adjustment in water treatment,” *Journal of Petroleum Research and Studies*, vol. 11, no. 3, pp. 19-27, Sep. 2021, doi: http://doi.org/10.52716/jprs.v11i3.530.

[2] M. Zainal-Abideen, A. Aris, F. Yusof, Z. Abdul-Majid, A. Selamat, and S. Omar, "Optimizing the coagulation process in a drinking water treatment plant–comparison between traditional and statistical experimental design jar tests," *Water science and Technology*, vol. 65, no. 3, pp. 496-503, 2012, doi: http://doi.org/10.2166/wst.2012.561.

[3] J. G. Gacu, C. E. F. Monjardin, R. G. T. Mangulabnan, G. C. E. Pugat, and J. G. Solmerin, "Artificial Intelligence (AI) in Surface Water Management: A Comprehensive Review of Methods, Applications, and Challenges," *Water*, vol. 17, no. 11, Art. no. 1707, 2025, doi https://doi.org/10.3390/w17111707.

[4] B. Senthil Rathi, P. Senthil Kumar, S. Sanjay, M. Prem Kumar, and G. Rangasamy, "Artificial intelligence integration in conventional wastewater treatment techniques: techno-economic evaluation, recent progress and its future direction," *International Journal of Environmental Science and Technology*, vol. 22, pp. 633-658, 2025, doi: http://doi.org/10.1007/s13762-024-05725-2.

[5] A. Alenezi and Y. Alabaiadly, "Artificial Intelligence Applications in Water Treatment and Desalination: A Comprehensive Review," *Water*, vol. 17, no. 8, Art. no. 1169, 2025, doi: http://doi.org/10.3390/w17081169.

[6] S. Chen, Y. Liu, I. Gorton, and A. Liu, "Performance prediction of component-based applications," *Journal of Systems and Software*, vol. 74, no. 1, pp. 35-43, 2005, doi: http://doi.org/10.1016/j.jss.2003.05.005.

[7] S. K. Kiangala and Z. Wang, "An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment," *Machine Learning with Applications*, vol. 4, Art. no. 100024, 2021, doi: http://doi.org/10.1016/j.mlwa.2021.100024.

[8] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data," *Atmosphere*, vol. 10, no. 7, Art. no. 373, 2019, doi: http://doi.org/10.3390/atmos10070373.

[9] K. Yetilmezsoy, B. Ozkaya, and M. Cakmakci, "Artificial intelligence-based prediction models for environmental engineering," *Neural Network World*, vol. 21, no. 3, 2011, doi: http://doi.org/10.14311/nnw.2011.21.012.

[10] Q. Zhang, Z. Li, S. Snowling, A. Siam, and W. El-Dakhakhni, "Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network," *Water Science and Technology*, vol. 80, no. 2, pp. 243-253, 2019, doi: http://doi.org/10.2166/wst.2019.263.

[11] S. Pan, Z. Zheng, Z. Guo, and H. Luo, "An optimized XGBoost method for predicting reservoir porosity using petrophysical logs," *Journal of Petroleum Science and Engineering*, vol. 208, Art. no. 109520, 2022, doi: http://doi.org/10.1016/j.petrol.2021.109520.

[12] D. S. Mahdi, E. A. Al-Khdheeawi, and Y. Yuan, "AI-Based Estimation of Poisson's Ratio for Carbonate Formations Using Drilling Parameters in a Southern Iraqi Oil Field," *Journal of Petroleum Research and Studies*, vol. 15, no. 2, pp. 44-59, 2025, doi: http://doi.org/10.52716/jprs.v15i2.1072.

[13] Cuantum Technologies LLC, *Feature Engineering for Modern Machine Learning with Scikit-Learn: Mastering Data Preparation and Transformation for Robust ML Models*. Birmingham, U.K.: Packt Publishing, 2025.

[14] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied soft computing*, vol. 90, Art. no. 106181, 2020, doi: http://doi.org/10.1016/j.asoc.2020.106181.

[15] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *Proc. 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*, Oslo, Norway, Jun. 5–9, 2022, doi: http://dx.doi.org/10.23967/eccomas.2022.155.