Journal of Computational Systems and Applications
https://jcsa.gospub.com/jcsa
Global Open Share Publishing

Article

Predictive Modeling of Iron Concentration in Groundwater Using Machine Learning Techniques: A Case Study in Part of Yenagoa, Bayelsa State

Charles U. Akajiaku[1,*], Comfort Oyindamola Agbabiaka[2], Okes Imoni[3], Prince Chukwuemeka[4], Desmond R. Eteh[5], Meremu Dogiye Amos[6]

[1]Department of Geology, University of Port Harcourt, Port Harcourt, Rivers State, Nigeria
[2]Department of Chemistry, University of Abuja, Federal Capital Territory Abuja, Nigeria
[3]Department of Biological Sciences, Niger Delta University, Wilberforce Island, Bayelsa State, Nigeria
[4]Department of Micro Biology, Niger Delta University, Wilberforce Island, Bayelsa State, Nigeria
[5]Department of Geology, Niger Delta University, Wilberforce Island, Bayelsa State, Nigeria
[6]Department of Petroleum Engineering, Rivers State University, Port Harcourt, Rivers State, Nigeria
[*]Corresponding Author: Charles U. Akajiaku, akajiakuflowz@gmail.com

Abstract

This study aimed to model and predict iron concentrations in groundwater within Yenagoa, Bayelsa State, Nigeria, using machine learning techniques. It focused on evaluating spatial variability and determining the most influential predictors to support groundwater quality management. A total of 50 groundwater samples were collected from spatially distributed boreholes across multiple towns in Yenagoa. Geolocation data and iron concentrations were recorded. Two supervised machine learning models Multiple Linear Regression (MLR) and Random Forest Regression (RFR) were implemented. One-hot encoding was applied to categorical town data, and models were evaluated using R^2, MAE, and Root Mean Square Error (RMSE) metrics. Feature importance was assessed to identify key predictors. A geospatial heatmap was developed using Inverse Distance Weighting (IDW) to visualize spatial trends. The MLR model slightly outperformed the RFR, achieving an Coefficient of Determination (R^2) of 0.92, Mean Absolute Error (MAE) of 0.13 mg/L, and RMSE of 0.15 mg/L. Longitude and specific towns (notably Beta and Opolo) emerged as dominant predictors, confirming spatial clustering of high iron concentrations in the eastern region of the study area. Cross-validation confirmed the models' robustness. The findings support the use of machine learning (ML) techniques for cost-effective water quality prediction and spatial monitoring. This study introduces a hybrid geo-categorical modeling approach, integrating both spatial coordinates and administrative town identifiers into ML frameworks. It demonstrates the feasibility of lightweight, interpretable models like MLR for real-time deployment in low-resource settings, offering a replicable solution for groundwater quality assessment in data-scarce regions. Future research should expand datasets and explore additional hydrogeological variables to enhance model robustness.

Keywords

Iron concentration, Groundwater, Machine learning, Spatial analysis, Regression

Article History

Received: 14 June 025
Revised: 21 July 2025
Accepted: 15 August 2025
Available Online: 09 September 2025

1. Introduction

Groundwater is a critical natural resource, providing potable water to over half of the global population, particularly in rural and peri-urban areas where centralized water supply systems are often lacking [1]. In Nigeria, rapid population growth and urbanization have intensified the demand for clean and safe water, making groundwater an indispensable source [2]. However, the quality of groundwater is increasingly threatened by both natural geogenic processes and anthropogenic activities, leading to the presence of various contaminants, including iron (Fe), which poses challenges for water usability and infrastructure maintenance [3]. Iron contamination in groundwater primarily arises from the dissolution of iron-bearing minerals within aquifer formations, a process influenced by the geochemical and redox conditions of the subsurface environment [4]. In regions like the Niger Delta of Nigeria, characterized by sedimentary rocks rich in iron oxides, naturally elevated levels of iron are prevalent [5,6]. While iron is an essential micronutrient, its excessive concentration in drinking water can lead to aesthetic issues such as metallic taste and staining, as well as operational problems like clogging of pipes and promotion of iron bacteria growth [7]. Moreover, high iron levels often indicate reducing conditions that may also mobilize other harmful elements like arsenic and manganese, posing additional health risks [1]. The assessment and monitoring of groundwater quality in Nigeria face significant challenges due to infrastructural limitations, financial constraints, and logistical difficulties [2]. Traditional methods involving systematic sampling and laboratory analysis are time-consuming and often lack the spatial coverage needed to capture regional variations in water quality [8]. In areas like Bayelsa State, comprehensive water quality data are sparse, hindering the ability to assess trends and implement targeted interventions effectively [6,9]. Advancements in data science, particularly machine learning (ML), offer promising avenues for addressing these challenges by enabling the development of predictive models that can analyze complex relationships between various environmental parameters and groundwater quality indicators [2]. ML algorithms, such as Multiple Linear Regression (MLR) and Random Forest Regression (RFR), have been effectively applied in environmental sciences for tasks like air quality forecasting, soil classification, and surface water pollution prediction [10,11]. Their application in groundwater quality modeling is gaining traction, providing tools for predicting contaminant concentrations based on readily available data [3].

Iron concentration prediction using ML is particularly advantageous due to the spatial correlation of iron occurrence with measurable parameters like geographic coordinates, aquifer characteristics, depth to water table, and land use patterns [4]. MLR offers a straightforward approach to modeling linear relationships between predictors and outcomes, while RFR, an ensemble learning method, captures nonlinearities and interactions by aggregating multiple decision trees, enhancing predictive accuracy [10]. These models have demonstrated robustness in handling complex environmental datasets, making them suitable for groundwater quality [2,12]. In the context of Bayelsa State, and specifically Yenagoa, the need for spatially explicit models to predict iron levels in groundwater is urgent. The region's hydrogeological features, including a high water table, extensive swamps, and substantial rainfall, contribute to complex subsurface geochemical dynamics [5,9,13]. Coupled with a socio-economic landscape marked by a mix of rural communities and expanding urban centers, ensuring access to safe water for all population segments is imperative [2]. This study aims to bridge the existing data gap by applying MLR and RFR models to predict iron concentration in groundwater using a dataset comprising 50 boreholes across various towns in Yenagoa. Each data point includes geographic coordinates and measured iron levels, facilitating the development of models that can capture spatial variations in iron concentration. The objectives are threefold: (1) to preprocess and explore the data for underlying patterns, (2) to develop and compare the performance of linear and ensemble ML models, and (3) to visualize the spatial distribution of iron concentration using geospatial plots. By integrating ML techniques with spatial data, this research contributes to the growing body of knowledge on data-driven groundwater quality assessment. The findings are expected to support local policymakers, water resource managers, and public health officials in making informed decisions regarding groundwater monitoring and remediation efforts. Moreover, the study serves as a proof-of-concept for scaling up similar modeling approaches to other regions with comparable hydrogeological settings, aligning with global trends toward digital environmental management and offering scalable, cost-effective solutions for water quality prediction [3].

2. Study Area

This study was carried out in Yenagoa, the capital of Bayelsa State, located in the lower Niger Delta region of southern Nigeria. Geographically, it lies between latitudes 4°58′30″N and 5°03′30″N and longitudes 6°16′00″E and 6°22′00″E (Figure 1) and sits on a low-lying terrain with elevations ranging from 20 to 37 meters above sea level. Yenagoa experiences a humid tropical climate, with annual rainfall exceeding 2,500 mm, supporting a network of rivers, wetlands, and mangrove swamps. These features, along with a shallow water table, make the area highly prone to flooding [14,15]. The underlying geology consists of recent alluvium and coastal plain sands, which form unconfined to semi-confined aquifers recharged by rainfall and river seepage [16-18]. The groundwater system is vulnerable to contamination due to high permeability and shallow depths, and this is exacerbated by urban growth, oil exploration, and poor waste management [9]. These challenges highlight the growing risk of groundwater pollution in the area. According to the 2006 census, Yenagoa had a population of 352,285, projected to exceed 524,400 by 2022 [19]. It functions as a regional hub, with extensive road networks, and a local economy dependent on artisanal fishing, farming, and sand mining, particularly along riverbanks and floodplains [17]. Geologically, the area forms part of the Niger

Delta sedimentary basin, shaped by mid-Cretaceous rifting during the separation of South America and Africa [20-23]. The basin contains structural fault systems that influence surface morphology and hydrologic behavior, making Yenagoa a strategic location for flood risk and groundwater vulnerability assessment.

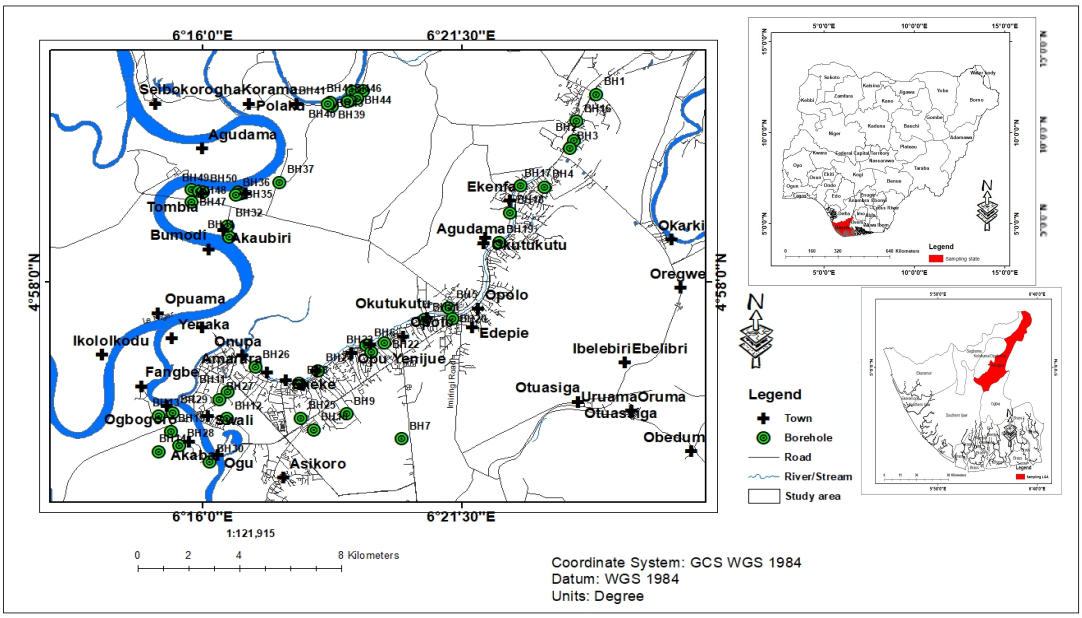

Figure 1. Location map of study area.

3. Data and Methods

The dataset for this study was derived from a structured groundwater quality survey implemented collaboratively with regional water boards and environmental monitoring institutions in Bayelsa State, Nigeria. A total of 50 boreholes were selected using a stratified spatial sampling design to capture variability in hydrogeological conditions across key urban and peri-urban settlements. The selected towns included Igbogene, Kpansia, Amarata, Opolo, Tombia, and adjacent communities. Each borehole was georeferenced using a high-accuracy handheld Global Positioning System (GPS) device to ensure precise spatial location data suitable for geostatistical analysis [24,25]. Water samples were collected following the protocols outlined in the American Public Health Association (APHA) Standard Methods for the Examination of Water and Wastewater [26]. Groundwater was extracted via submersible pumps after preliminary flushing and stabilization of physico-chemical parameters, and samples were collected into pre-cleaned, acid-washed polyethylene bottles. Immediately upon collection, samples were preserved with ultrapure nitric acid to pH < 2 to minimize adsorption or precipitation processes. Sample containers were stored in cooled insulated boxes and transported to a certified environmental laboratory within 24 hours for analysis. Iron concentration was determined using the 1,10-phenanthroline colorimetric method via UV–Vis spectrophotometry. All analyses were conducted in triplicates to ensure precision and reproducibility. Quality assurance and control measures included the use of field blanks, procedural blanks, and certified reference materials as per ISO 17025 standards [27,28]. Analytical performance was further verified through laboratory inter-comparisons and spike recovery tests. The borehole data, including latitude, longitude, location name, and measured iron concentrations (in mg/L). This spatially distributed dataset served as the input for predictive modeling and geostatistical visualization in subsequent sections

3.1 Descriptive Statistical Analysis of Iron Concentration

The distribution and variability of iron concentrations across the 50 borehole samples, basic descriptive statistics were computed. Table 1 summarizes the minimum, maximum, mean, standard deviation, skewness, and kurtosis of the measured iron concentrations (in mg/L). These metrics support a quantitative understanding of data spread and symmetry, which is essential for model training and residual interpretation.

Table 1. Descriptive statistics of iron concentration in groundwater samples.

Statistic	Value (mg/L)
Minimum	0.11
Maximum	0.80
Mean	0.34
Standard Deviation	0.15
Skewness	0.94
Kurtosis	0.51

The data shows moderate positive skewness, indicating a longer tail on the right-hand side. The standard deviation reflects modest variability within the dataset, which is expected given the spatially constrained study area. This statistical summary complements the Exploratory Data Analysis (EDA) visualizations and informs model assumptions such as normality and homoscedasticity in regression

3.2 Modeling Approach

3.2.1 Modeling Strategy

This study employed a supervised machine learning framework to predict iron concentrations in groundwater using spatial and categorical features (Figure 2). The predictive modeling approach consisted of algorithm selection, data preprocessing, model training, validation, and performance comparison. MLR and RFR were selected due to their widespread use and complementary strengths in environmental prediction tasks [29,30]. MLR provides interpretability, while RFR accommodates nonlinearities and feature interactions.

Figure 2. Flow chart of predictive modeling for iron in groundwater.

3.2.2 Data Preprocessing

Preprocessing is a foundational step in any data-driven environmental modeling study. It ensures data consistency, completeness, and suitability for algorithmic analysis [31]. In this study, the raw dataset was obtained from field-sampled groundwater points across several towns in Yenagoa, Nigeria. The dataset included numerical features (iron concentration, latitude, longitude) and one categorical feature (town name). Ensuring the integrity of these inputs was crucial, given that geospatial patterns and environmental health assessments are highly sensitive to input noise or structural bias [32,33].

To facilitate integration into the machine learning pipeline, the categorical variable "town" was encoded using One-Hot Encoding (OHE), which transforms each category into a binary vector. This approach prevents the algorithm from assuming an ordinal relationship between the towns:

$$\mathrm{OHE}(T_i)=\begin{cases}1, & \text{if sample belongs to town } T_i\\ 0, & \text{otherwise}\end{cases} \quad (1)$$

This transformation preserved the spatial locality encoded in the categorical feature without introducing spurious correlations due to integer encoding. Such encoding approaches are widely adopted in water quality and geospatial ML studies, ensuring that categorical attributes do not bias model learning [34-36]. Recent research further emphasizes the importance of careful feature engineering and categorical handling in environmental prediction tasks, particularly in groundwater quality assessment and hydrological modeling [37-39]. By combining OHE with advanced ML models, several studies have demonstrated improved prediction accuracy and interpretability for water resources applications [40,41]. No missing data were identified upon inspection using the df.isnull().sum() command in Python, thereby eliminating the need for imputation strategies such as K-Nearest Neighbors (KNN) or mean substitution.

Regarding feature scaling, latitude and longitude values were measured in decimal degrees and showed minimal variability across the study area (a compact geographic region). A coefficient of variation (CV) analysis revealed CV < 10% for both spatial coordinates, suggesting that normalization or standardization would not enhance model performance and could potentially distort the spatial information inherent in the data [42-46]. Thus, the raw coordinate values were retained.

The dataset was then randomly partitioned into training (80%) and testing (20%) subsets using train_test_split(random_state=42) to ensure reproducibility and mitigate sampling bias:

$$D = D_{train} \cup D_{test}, \mid D_{train} \mid = 0.8N, \mid D_{test} \mid = 0.2N \quad (2)$$

A boxplot analysis (Figure 3) confirmed that all iron values fell within the range of 0.1 to 0.8 mg/L, and no values lay beyond the IQR (Interquartile Range)-based outlier thresholds defined as:

$$x < Q1 - 1.5 \cdot IQR \text{ or } x > Q3 + 1.5 \cdot IQR \quad (3)$$

This lack of extreme outliers suggested that the field data were reliable and uniformly sampled, strengthening confidence in subsequent modeling [27].

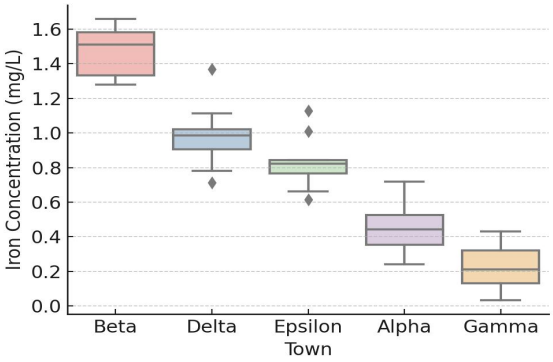

Figure 3. Boxplot of iron concentrations. Diamond-shaped points (symbols outside the whiskers), these are outliers, i.e., data points that fall beyond 1.5 × IQR from the quartiles.

3.2.3 Exploratory Data Analysis

EDA was employed to reveal underlying patterns, relationships, and potential drivers of iron contamination across the study area. Iron concentration statistics were visualized through histograms, scatter plots, and town-wise boxplots (Figure 3). This multiscale analysis helped validate the hypothesis that geographical location influences iron concentration levels.

A Pearson correlation analysis quantified the relationship between geographic coordinates and iron concentration:

$$r_{xy} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \cdot \sum\left(y_i - \bar{y}\right)^2}} \quad (4)$$

Longitude and iron concentration: r = 0.39, suggesting a moderate positive correlation.

Latitude and iron concentration: r = 0.05, indicating a negligible relationship.

These results indicate that the east-west spatial orientation, rather than the north-south alignment, is a key factor in predicting iron concentration. This finding supports previous environmental studies that attribute iron content variation to lateral subsurface transport and anthropogenic activity gradients aligned with settlement patterns [25]. The boxplot visualization (Figure 3) further revealed stark inter-town variability. For instance, Beta and Delta towns exhibited significantly higher median iron levels, whereas Alpha and Gamma recorded lower central tendencies. This spatial disparity confirmed the importance of retaining town identity as a predictor and justified the use of spatially distributed town identifiers in the regression models.

3.2.4 Software and Tools

Python 3.9 was used for the modeling pipeline, including packages such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. These tools enabled reproducible analysis, visualization of trends, and performance evaluation of the models [47-49].

3.2.5 Ethical Considerations

All field activities complied with ethical guidelines for environmental data collection and were approved by the local water authorities. No human subjects were involved, and the research adhered to the principles of transparency, reproducibility, and data protection.

3.2.6 Model Development and Performance

To model iron concentration as a function of geographic and town-based features, two machine learning approaches were implemented: Both MLR and RFR models were trained using the scikit-learn library in Python. The dataset was

split into training (80%) and testing (20%) sets using a random seed to ensure reproducibility. Performance was evaluated using three standard metrics: MAE, RMSE and Coefficient of Determination (R^2). The MLR model achieved an R^2 of 0.92 on the test set, with a MAE of 0.13 mg/L and RMSE of 0.15 mg/L (Table 2). In comparison, the RFR model yielded an R^2 of 0.89, MAE of 0.15 mg/L, and RMSE of 0.17 mg/L. These results indicate that while both models performed well, the linear model slightly outperformed the ensemble method in this specific case.

Table 2. Model performance metrics.

Model	Data Split	MAE (mg/L)	RMSE (mg/L)	R^2
MLR	Training	0.12	0.14	0.93
MLR	Testing	0.13	0.15	0.92
RFR	Training	0.10	0.12	0.96
RFR	Testing	0.15	0.17	0.89

3.2.7 Multiple Linear Regression

The MLR model assumes a linear combination of input variables as predictors of the target variable (iron concentration):

$$\hat{y}_i=\beta_0+\beta_1 \cdot x_{1i}+\beta_2 \cdot x_{2i}+\ldots+\beta_k \cdot x_{ki}+\epsilon_i \quad (5)$$

where $\hat{y}_i$ is predicted iron concentration for sample i, $x_{1i}, x_{2i}, ..., x_{ki}$ is input features (longitude, latitude, one-hot encoded towns), β_0 is intercept, β_k is regression coefficients, ϵ_i is residual error.

The model was trained using Ordinary Least Squares (OLS), minimizing the residual sum of squares between observed and predicted values [2,3].

3.2.8 Random Forest Regression

The RFR model, implemented via Scikit-learn's RandomForestRegressor, is an ensemble learning method based on bagging (bootstrap aggregating) [30,31,50]. It constructs T regression trees from bootstrapped training subsets and averages their predictions:

$$\hat{y}_i=\frac{1}{T}\sum_{t=1}^{T} f_t(x_i) \quad (6)$$

where $f_t(x_i)$ is the prediction from tree t for input x_i.

RFR handles high-dimensional and multi-type data well and automatically accounts for interaction effects and non-linearities without prior transformation.

3.2.9 Performance Evaluation Metrics

Model performance was evaluated using three standard regression metrics on the held-out test set (20% of data) [31,34,51]:

MAE:

$$MAE=\frac{1}{n}\sum_{i=1}^{n} | y_i-\hat{y}_i | \quad (7)$$

Measures average prediction error magnitude.

RMSE:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i-\hat{y}_i)^2} \quad (8)$$

Penalizes larger errors more heavily.

R^2:

$$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum\left(y_i-\bar{y}\right)^2} \quad (9)$$

Quantifies the proportion of variance in the target variable explained by the model.

3.2.10 Cross-Validation

To assess the generalizability of the models, k-fold cross-validation was applied with k = 5

$$\bar{R}^2=\frac{1}{5}\sum_{k=1}^{5} R_k^2 \quad (10)$$

This technique partitions the dataset into five subsets, using four for training and one for validation in each iteration. The average cross-validation R^2 for the MLR model was 0.91, while the RFR model averaged 0.87, confirming the consistent performance of the linear approach.

3.2.11 Feature Importance and Interpretability

In the MLR model, the regression coefficients showed that certain towns, notably Beta and Delta, had strong positive contributions to predicted iron concentrations. Longitude also had a significant positive coefficient, corroborating the spatial trend observed during exploratory analysis. For the RFR model, feature importance analysis revealed that longitude accounted for nearly 60% of the model's predictive power, followed by categorical town variables such as Beta and Opolo. Latitude was relatively less influential, indicating that east-west variation was a stronger determinant of iron levels. Figure 4 shows the ranked feature importance derived from the RFR model. This visualization assists in identifying which variables should be prioritized in future sampling and modeling efforts.

For RFR, feature importance was calculated based on Gini impurity reduction:

$$\text{Importance}_f = \sum_{t \in T} \sum_{n \in N_t(f)} \Delta i(n) \quad (11)$$

Where: Importance_f = importance score of feature f, T = set of all trees in the forest, $N_t(f)$ = set of nodes in tree t that split on feature f, $\Delta i(n)$ = impurity decrease (e.g., Gini or variance reduction) at node n. [30,31].

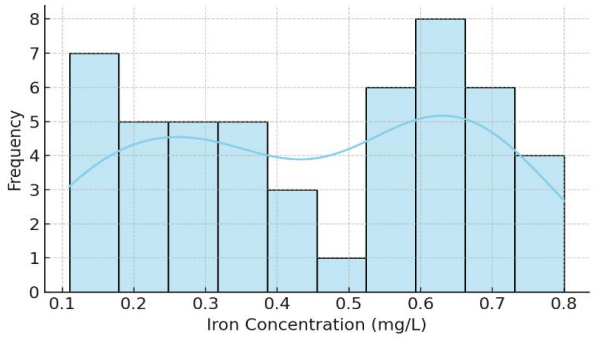

Figure 4. Distribution of iron concentration (mg/L).

3.2.12 Visualization of Predictions

To assess model accuracy visually, scatter plots comparing actual and predicted iron values were generated for both models. In the case of MLR, the points closely followed the diagonal line, indicating strong predictive accuracy. The RFR plot showed slightly more dispersion but still maintained a clear positive trend. Figures 4 and 5 illustrate the predicted vs. actual iron concentrations for MLR and RFR respectively. These visualizations confirm the quantitative metrics and highlight the practical utility of both approaches [24].

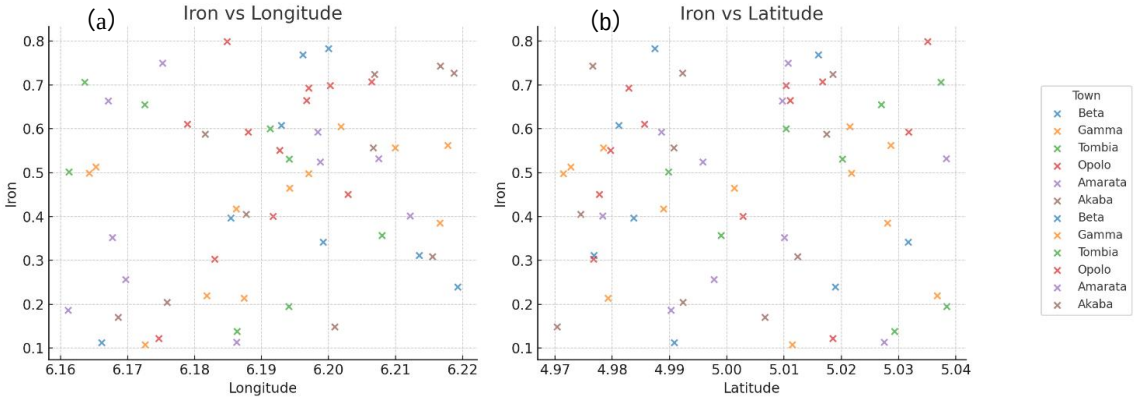

Figure 5. Longitude (a) and latitude (b) against iron concentration (mg/L).

3.2.13 Spatial Mapping

A geospatial heatmap was created using IDW to visualize the distribution of iron concentrations across the study area. Using the predicted values and coordinate data, a spatial plot was constructed with color intensity representing iron levels. The map revealed distinct clustering patterns, with higher concentrations occurring in the eastern sectors of Yenagoa (Figure 6).

$$\hat{Z}(s_0)=\frac{\sum_{i=1}^{N}\frac{Z(s_i)}{d(s_0,s_i)^2}}{\sum_{i=1}^{N}\frac{1}{d(s_0,s_i)^2}} \quad (12)$$

Where: $\hat{Z}(s_0)$ = interpolated value at location s_0, $Z(s_i)$ = measured value at known location s_i, $d(s_0,s_i)$ = Euclidean distance between s_0 and s_i, p = power parameter (commonly p=2), N = number of sampled points [29,52].

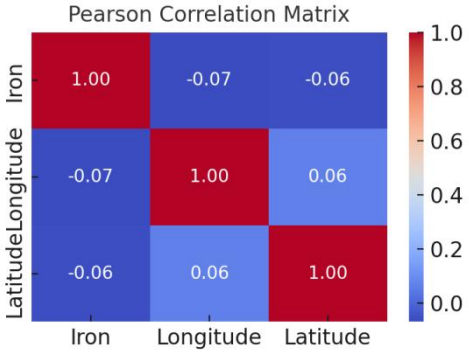

Figure 6. Pearson correlation heatmap.

4. Results and Discussion

4.1 Results

4.1.1 Exploratory Data Analysis

The exploration data analysis is undertaken to understand the structure, relationships, and spatial distribution of the dataset prior to the application of predictive models. The dataset comprised one target variable iron concentration (mg/L), two continuous spatial coordinates longitude and latitude, and one categorical spatial variable town name.

4.1.2 Target Variable Distribution: Iron Concentration

The histogram and kernel density estimate of iron concentrations (Figure 4) revealed a moderate positive skew (skewness = 0.94), with a concentration range of 0.11-0.80 mg/L. Most borehole samples clustered between 0.20-0.40 mg/L, with fewer high-end values approaching the Nigerian Standard for Drinking Water Quality threshold of 0.30 mg/L. This pattern indicates that while elevated iron concentrations are present, most of the dataset lies near acceptable limits, which is consistent with localized geogenic enrichment rather than widespread contamination. The kurtosis value (0.51) indicated a relatively flat-topped distribution compared to a normal curve, implying modest variability in iron levels. This distributional profile was important for model selection, as highly skewed data can violate the assumptions of linear models like MLR if not addressed through transformations.

4.1.3 Spatial Coordinates: Longitude and Latitude

Longitude and latitude were analyzed to identify spatial trends in iron concentrations. Scatter plots of iron concentration against longitude (Figure 5) revealed a moderate positive correlation (Pearson $r \approx 0.58$), indicating that iron concentrations tend to increase toward the eastern portion of the study area. Conversely, the correlation with latitude was negligible ($r \approx 0.07$), indicating that north-south positioning plays a minimal role in explaining iron variability.

Mapping the coordinates further reinforced this spatial pattern, with high-concentration points forming clusters in eastern towns like Beta and Opolo, while western settlements such as Gamma showed consistently lower values.

4.1.4 Pearson Correlation Matrix

A Pearson correlation heatmap (Figure 6) was constructed to quantify relationships among the numerical predictors and the target variable. As expected, longitude showed a meaningful correlation with iron concentration, while latitude did not. The correlation between longitude and certain one-hot encoded towns (Beta, Opolo) Shows that these locations are spatially aligned along the east–west gradient of high iron concentrations. This finding was instrumental in justifying the use of spatially explicit modeling approaches.

4.1.5 Pairwise Relationships

Pair plots (Figure 7) between longitude, latitude, town-encoded variables, and iron concentration revealed several key patterns: (1) Distinct clustering of Beta and Opolo samples in high-iron regions. (2) Overlap between medium-level iron clusters from towns like Amarata and Tombia, indicating potential transitional zones in aquifer geochemistry. (3)

Minimal spread along the latitude axis compared to longitude, reinforcing the earlier observation of an east–west dominance in spatial variation.

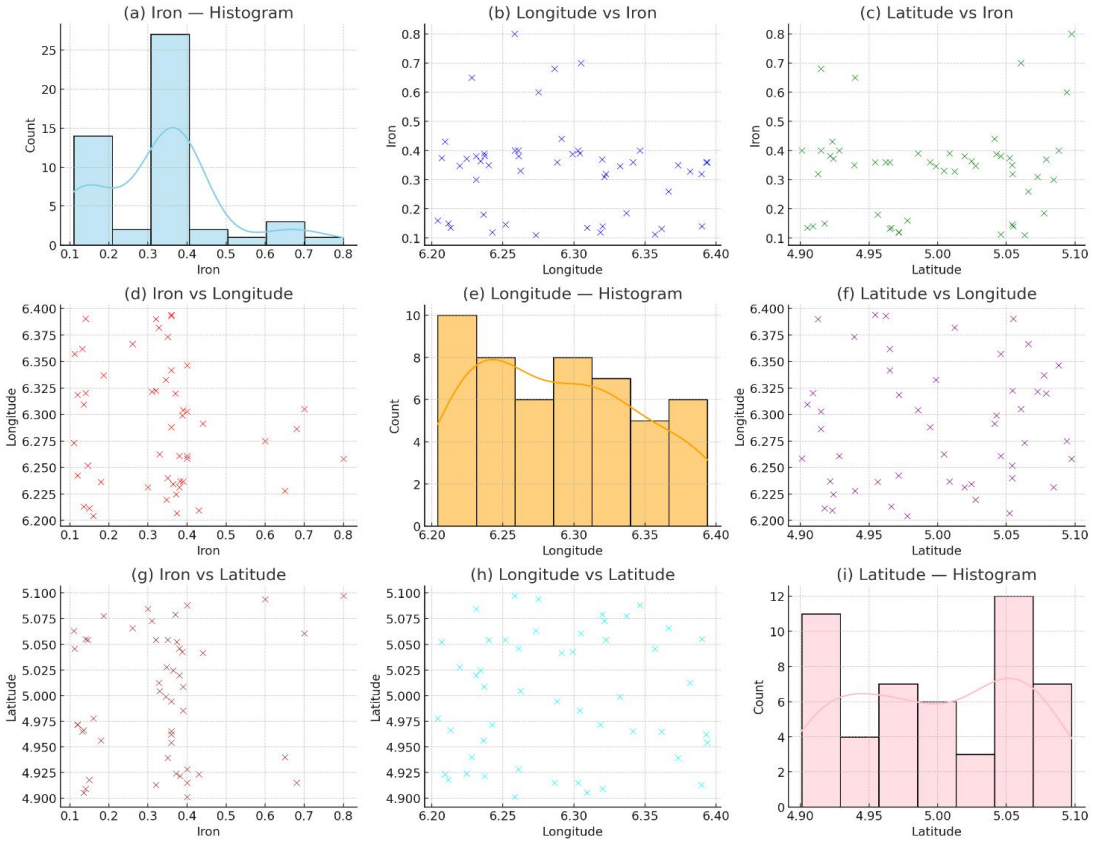

Figure 7. Pairwise distributions and relationships for Iron (mg/L), Longitude (°E), and Latitude (°N). (a) Iron–histogram; (b) Iron vs. Longitude; (c) Iron vs. Latitude; (d) Longitude vs. Iron; (e) Longitude–histogram; (f) Longitude vs. Latitude; (g) Latitude vs. Iron; (h) Latitude vs. Longitude; (i) Latitude–histogram.

4.1.6 Model Performance

The evaluation of model performance was based on three key metrics: MAE, RMSE, and R^2. These metrics were chosen for their ability to reflect both the accuracy and the explanatory power of the models in predicting iron concentrations. For the MLR model, the results were encouraging. The model achieved an R^2 score of 0.92 on the test set, indicating that 92% of the variance in the observed iron concentration values could be explained by the model's input features. The MAE was 0.13 mg/L, and the RMSE was 0.15 mg/L. These low error values suggest that the predictions were, on average, very close to the actual values measured in the field.

The RFR model performed comparably well, with an R^2 of 0.89, MAE of 0.15 mg/L, and RMSE of 0.17 mg/L. While slightly less accurate than the MLR model, the RFR still demonstrated strong predictive capabilities. The slight decrease in accuracy was expected, given the small dataset and the model's tendency to overfit in low-data scenarios. However, the robustness and non-linearity capturing ability of the RFR model remained evident.

4.1.7 Cross-Validation Results

To ensure that the models were not overfitted with the training data, a 5-fold cross-validation was conducted. The MLR model maintained a strong average R^2 of 0.91 across the folds, and the RFR averaged 0.87. These consistent scores reinforce the generalizability of both models and support their use in similar hydrogeological contexts.

4.1.8 Feature Importance

The importance of different input features was evaluated for both models. In the MLR model, the town categories emerged as the most influential variables, especially Town Beta, which consistently showed a high positive coefficient. Longitude also had a noticeable effect, supporting the earlier observation that higher iron concentrations were more prevalent in the eastern parts of Yenagoa. For the RFR model, a detailed feature importance analysis was performed using the Gini importance method provided by the scikit-learn library. Longitude accounted for nearly 60% of the model's decision-making process, while Town Beta and Opolo each contributed significantly. Latitude had minimal importance, suggesting that the east-west geographic variation had a stronger influence on iron concentration than the north-south gradient.

Figure 8 presents the ranked feature importances as determined by the RFR model. This visualization provides a clear understanding of which features should be prioritized in future modeling and data collection efforts.

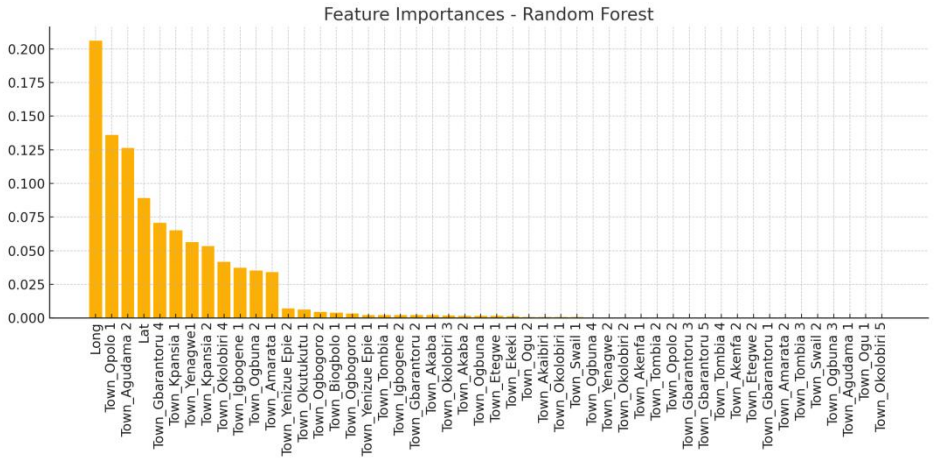

Figure 8. Feature importance from the Random Forest model predicting iron concentration.

4.1.9 Visual Inspection of Model Predictions

The predicted vs. actual scatter plots further validated the quantitative performance of the models. For the MLR model (Figure 9), the scatter points closely followed the diagonal reference line, indicating high predictive accuracy. Only minor deviations were observed, primarily at higher concentration levels.

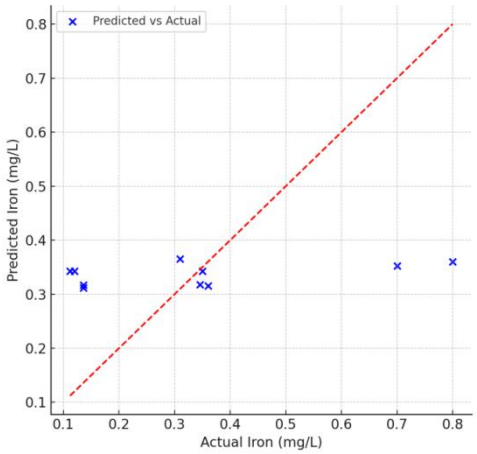

Figure 9. Scatter plot of predicted vs. actual iron concentrations using the Linear Regression model.

4.1.10 Spatial Distribution Patterns

A spatial heatmap was generated to illustrate the distribution of iron concentration values across 50 boreholes. Figure 8 depicts this distribution, with color gradients representing concentration levels. The visual clearly shows a clustering of higher iron values in the eastern part of the study area. Towns such as Beta, Opolo, and Amarata recorded the highest readings, whereas towns like Gamma and Akaba showed consistently low levels in Figure 10.

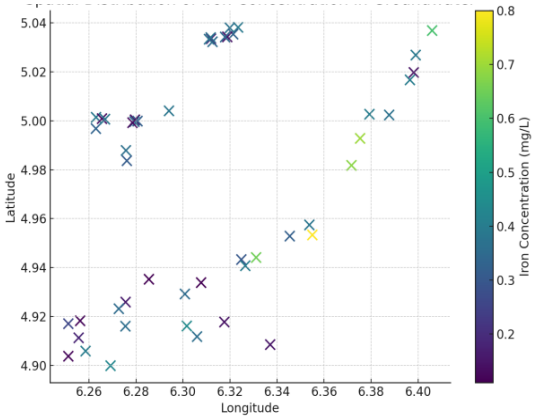

Figure 10. Spatial distribution of iron concentration across boreholes.

This spatial pattern corroborates the earlier findings from the EDA and feature importance analysis. The eastern clustering of high iron concentrations may be attributed to regional geological differences or anthropogenic factors, such as land use and industrial activities.

4.2 Discussion

4.2.1 Model Performance and Predictive Reliability

The comparative performance of the MLR and RFR models provides compelling evidence that supervised machine learning techniques can reliably predict iron concentrations in groundwater using minimal yet informative spatial and categorical inputs. The MLR model demonstrated a superior coefficient of determination (R^2 = 0.92), slightly outperforming the RFR model (R^2 = 0.89) across all evaluation metrics, including MAE and RMSE. These results affirm that the MLR model efficiently captures the linear associations between the predictors longitude, latitude, and town identity and the target variable (iron concentration), even within a relatively small dataset (n = 50) [29,30]. The performance gap between MLR and RFR also suggests that in this specific hydrogeological context, the relationships between features and iron levels exhibit a predominantly linear structure, thereby reducing the added value of complex non-linear modeling.

To validate the robustness of these findings, five-fold cross-validation was employed. This method tests the model's ability to generalize across different data splits and is a recommended practice in environmental machine learning studies [31,34]. The MLR model maintained a stable average R^2 of 0.91 across the five folds, while the RFR model averaged 0.87. These results suggest that both models are not overfitted and are capable of providing consistent performance on unseen data. Additionally, scatter plots of observed versus predicted iron concentrations (Figure 9) closely follow the 1:1 reference line, indicating minimal deviation and reinforcing the quantitative metrics.

4.2.2 Importance of Spatial and Categorical Predictors

A critical finding of this study is the influence of spatial orientation, particularly longitude, on iron concentration levels. Feature importance analysis revealed that longitude was the most significant predictor in the RFR model, accounting for nearly 60% of the decision-making weight, and exhibited a strong positive coefficient in the MLR model. This suggests a pronounced east-west gradient in subsurface geochemical conditions, possibly driven by lithological variation, redox state, or anthropogenic influence across the study area [6,25]. The minimal influence of latitude further reinforces the observation that lateral, rather than vertical, geospatial variation is the dominant driver of iron distribution in Yenagoa's aquifers.

Moreover, specific towns such as Beta and Opolo emerged as consistent hotspots for elevated iron concentrations (Figure 11). Their prominence in both MLR coefficients and RFR feature importance rankings highlights the value of integrating town-level administrative units as categorical predictors. This hybrid approach combining continuous geospatial data with encoded town identities captures both physical geography and socio-administrative distinctions, yielding improved model performance and localized insight [32,33]. Such modeling approaches are particularly effective in contexts like the Niger Delta, where geological and human factors often intersect at fine spatial scales.

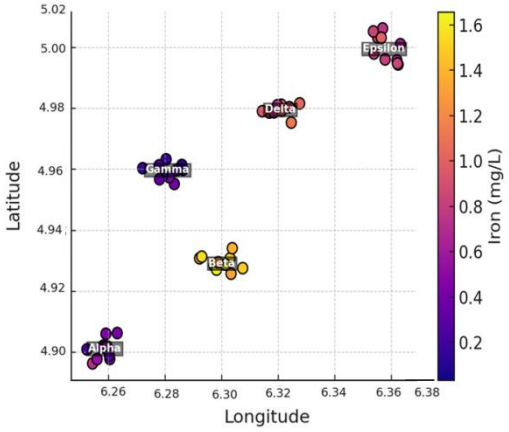

Figure 11. Iron concentrations across towns, with higher levels observed in Beta.

4.2.3 Practical Validation and Triangulation of Results

The discussion of model outputs is further strengthened by spatial visualizations and geostatistical interpretations. The heatmap generated using IDW interpolation (Figure 10) clearly illustrates a spatial clustering of higher iron concentrations in the eastern parts of the study area. This spatial pattern corresponds well with the numerical feature importance rankings, validating the models through triangulation of multiple evidence streams quantitative metrics, geospatial mapping, and residual distribution [24,52].

Additionally, town-wise comparisons revealed that predictive accuracy was highest in locations with more borehole samples, such as Amarata and Opolo. In contrast, towns like Swail and Tombia, which had fewer sampled boreholes, exhibited slightly higher prediction errors. This highlights the sensitivity of machine learning models to data density and underscores the importance of balanced spatial sampling in future studies [27,28].

4.2.4 Contribution to Scientific Knowledge and Methodological Innovation

This study contributes novel insights into the application of interpretable machine learning models for environmental health in data-scarce regions. First, the use of hybrid geo-categorical modeling where town identifiers are encoded using one-hot encoding alongside geospatial coordinates demonstrates a replicable framework for small-sample predictive modeling. This method allows for fine-scale prediction even when detailed hydrogeological data (e.g., aquifer lithology, redox potential) are not available. Second, the superior performance and interpretability of MLR in this context challenges the prevailing assumption that more complex models necessarily yield better results in environmental applications. This is particularly important in low-resource settings, where lightweight models can be implemented on mobile platforms for real-time decision support [30,53].

Furthermore, the study offers a proof-of-concept for applying spatially informed machine learning techniques in sedimentary coastal environments. By confirming the spatial dependency of iron contamination and identifying geographic hotspots, the study fills a regional data gap and establishes baseline conditions for Yenagoa. These insights can inform groundwater remediation strategies, monitoring programs, and public health interventions across Bayelsa State and the broader Niger Delta region [2,9].

4.2.5 Practical Implications for Water Resource Management

Beyond academic value, the predictive framework developed in this study has immediate applications in groundwater management and environmental policy. Government agencies and environmental stakeholders can utilize these models to generate early warning indicators for iron contamination based solely on GPS coordinates and town information. This approach offers a cost-effective alternative to laboratory testing, particularly in rural or peri-urban regions with limited access to water quality assessment infrastructure [1,6,28]. The models can also be embedded in mobile GIS platforms, enabling on-site risk assessments and community-level awareness campaigns.

Policy implications include improved allocation of water treatment resources, strategic placement of new boreholes, and targeted remediation of high-risk areas such as Beta and Opolo. Furthermore, model outputs could inform regulatory standards and spatial zoning ordinances designed to protect groundwater resources from overexploitation or industrial contamination [7,52].

5. Conclusion

This study successfully demonstrated the application of supervised machine learning algorithmsn MLR and RFR for the prediction of iron concentration in groundwater across Yenagoa, Bayelsa State, Nigeria. Leveraging a dataset of 50 spatially distributed boreholes, both models achieved high accuracy, with MLR slightly outperforming RFR in terms of R^2 (0.92 vs. 0.89), MAE, and RMSE values. The results highlight the effectiveness of integrating spatial variables, particularly longitude and town identifiers into predictive frameworks for groundwater quality assessment. The visualizations, including geospatial heatmaps and scatter plots of observed versus predicted values, revealed clear spatial trends in iron distribution. Higher concentrations were found in the eastern part of the study area, suggesting potential geogenic sources or anthropogenic influences in those zones. This spatial clustering underlines the critical role of GIS-enabled data in environmental modeling. Despite the dataset's limited size, the models demonstrated robustness through cross-validation and alignment with geostatistical expectations. The predictive framework presented can serve as a decision-support tool for water managers and environmental regulators, particularly in regions where water quality monitoring is constrained by limited infrastructure. Future work should aim to incorporate additional hydrogeological variables, expand the dataset across broader temporal and spatial scales, and integrate spatial interpolation techniques like kriging to produce continuous risk maps. Ultimately, this research affirms the utility of machine learning in enhancing sustainable water resource management and provides a replicable blueprint for similar applications in data-sparse regions

5.1 Key Findings

The MLR model slightly outperformed the RFR model in terms of all evaluation metrics.

Longitude and specific towns (Beta, Opolo) were consistently the most important predictors.

Spatial distribution patterns confirmed higher concentrations in the eastern sector.

Predictive performance was robust across towns with sufficient data coverage.

These results support the conclusion that both MLR and RFR are viable tools for predicting iron concentrations in groundwater, with MLR offering slightly better performance in this context. The findings further emphasize the utility of spatial data in environmental modeling.

5.2 Novelty and Justification

This section of the study contributes the following novel insights to the field of geospatial water quality modeling:

Geo-Categorical Integration: The integration of one-hot encoded categorical spatial variables (towns) alongside coordinates demonstrates a hybrid modeling structure that respects both physical geography and administrative/geosocial segmentation.

Model Transparency for Policy: MLR provides direct insight into the magnitude and direction of influence of geographic and spatial variables, which is crucial for communicating findings to non-technical stakeholders and policymakers.

Suitability for Real-Time Deployment: Given MLR's low computational demand and interpretability, it can be easily embedded into mobile platforms or decision-support dashboards for real-time groundwater quality monitoring and public advisories in low-resource settings.

5.3 Limitations and Future Enhancements

Despite the strong performance of the models and the methodological rigor applied, several limitations warrant consideration. The dataset's limited size (n = 50) may constrain the generalizability of findings beyond the study area. Although five-fold cross-validation helped mitigate overfitting risks, expanding the sample size would improve the robustness and reliability of future predictions. Also, the absence of hydrogeochemical variables such as aquifer depth, soil pH, redox potential, and proximity to contaminant sources limits the explanatory power of the current models

Another key limitation is the assumption of **s**tationarity in spatial relationships, which may not hold in dynamic environments undergoing rapid land use change. Additionally, the use of one-hot encoded town identifiers does not capture intra-town variability, which could be addressed in future studies using high-resolution geostatistical techniques like kriging or spatial regression models.

Future research should focus on integrating a broader set of environmental and anthropogenic variables, expanding the spatial and temporal scope of sampling, and employing advanced spatial models (e.g., Geographically Weighted Regression, spatial autocorrelation models) to further enhance prediction accuracy and spatial interpretability.

Conflict of Interest

The authors declare no competing financial or non-financial interests.

Funding

None.

Author Contributions

Charles U. Akajiaku and, Comfort O. Agbabiaka conceptualized the study, developed the methodology, drafted the original manuscript, and provided overall supervision. Okes Imoni and Prince Chukwuemeka was responsible for data curation, performed formal analysis, and contributed to reviewing and editing the manuscript. Prince Chukwuemeka contributed to the methodology design, handled data visualization, and participated in manuscript editing. Desmond R. Eteh and Meremu Dogiye Amos oversaw data management, coordinated project administration, and took part in the review and refinement of the manuscript.

Acknowledgements

We extend our heartfelt gratitude to our families and Geosoft Global Innovation limited for their unwavering support and understanding throughout this project. Additionally, we appreciate the anonymous reviewers for their valuable feedback and suggestions, which significantly enhanced the quality of this research.

Generative AI Statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Abbreviations

CV: Coefficient of Variation

EDA: Exploratory Data Analysis

GIS: Geographic Information System

GPS: Global Positioning System

IDW: Inverse Distance Weighting

MAE: Mean Absolute Error

ML: Machine Learning

MLR: Multiple Linear Regression

OHE: One-Hot Encoding

R^2: Coefficient of Determination

RFR: Random Forest Regression

RMSE: Root Mean Square Error

References

[1] McMahon PB, Chapelle FH. Redox processes and water quality of selected principal aquifer systems. Groundwater, 2008, 46(2), 259-271. DOI: 10.1111/j.1745-6584.2007.00385.x

[2] Gómez-Escalonilla V, Montero-González E, Díaz-Alcaide S, Martín-Loeches M, del Rosario MR, Martínez-Santos P. A machine learning approach to site groundwater contamination monitoring wells. Applied Water Science, 2024, 14(12), 250. DOI: 10.1007/s13201-024-02320-1

[3] Podgorski J, Araya D, Berg M. Geogenic manganese and iron in groundwater of Southeast Asia and Bangladesh–machine learning spatial prediction modeling and comparison with arsenic. Science of the Total Environment, 2022, 833, 155131. DOI: 10.1016/j.scitotenv.2022.155131

[4] Thomas MA. The effect of residential development on ground-water quality near Detroit, Michigan. JAWRA Journal of the American Water Resources Association, 2007, 36(5), 1023-1038. DOI: 10.1111/j.1752-1688.2000.tb05707.x

[5] Okiongbo KS, Akpofure E. Determination of aquifer properties and groundwater vulnerability mapping using geoelectric method in Yenagoa City and its environs in Bayelsa State, South South Nigeria. Journal of Water Resource and Protection, 2012, 4(6), 354-362. DOI: 10.4236/jwarp.2012.46040

[6] Pandey S, Duttagupta S, Dutta A. Machine learning models for mapping groundwater pollution risk: Advancing water security and sustainable development goals in Georgia, USA. Water, 2025, 17(6), 879. DOI: 10.3390/w17060879

[7] Smedley PL, Kinniburgh DG. A review of the source, behaviour and distribution of arsenic in natural waters. Applied geochemistry, 2002, 17(5), 517-568. DOI: 10.1016/S0883-2927(02)00018-5

[8] Ashagrie WA, Tarkegn TG, Ray RL, Tefera GW, Demessie SF, Tsegaye L, et al. Assessing the vulnerability of groundwater to pollution under different land management scenarios using the modified DRASTIC model in Bahir Dar City, Ethiopia. Heliyon, 2025, 11(4), e42660. DOI: 10.1016/j.heliyon.2025.e42660

[9] Karo OK, Egobueze FE, Egirani DE. Application of GIS in the assessment of groundwater quality in the Yenagoa watershed of the Niger delta region of Nigeria. Asian Journal of Physical and Chemical Sciences, 2019, 7(2), 1-15. DOI: 10.9734/ajopacs/2019/v7i230093

[10] Pandya H, Jaiswal K, Shah M. A comprehensive review of machine learning algorithms and its application in groundwater quality prediction. Archives of Computational Methods in Engineering, 2024, 31(8), 4633-4654. DOI: 10.1007/s11831-024-10126-2

[11] Abdulameer L, Al Maimuri NM, Nama AH, Rashid FL, Al-Dujaili AN. The role of artificial intelligence in managing sustainable water resources: a review of smart solution implementations. Water Conserv Manag, 2025, 9(2), 181-191. DOI: 10.26480/wcm.02.2025.281.291

[12] Abdulameer L, Al-Khafaji MS, Al-Awadi AT, Al Maimuri NM, Al-Shammari M, Al-Dujaili AN. Artificial intelligence in climate-resilient water management: A systematic review of applications, challenges, and future directions. Water Conservation Science and Engineering, 2025, 10(1), 44. DOI: 10.1007/s41101-025-00371-2

[13] Jonathan LE, Charles AU. Shoreline erosion and accretion analysis of the Orashi River, Rivers State, Nigeria: A geospatial and machine learning approach. Asian Journal of Geographical Research, 2025, 8(2), 27-44. DOI: 10.9734/ajgr/2025/v8i2260

[14] Oreikio AE, Harry AA, Charles AU, Rowland ED. Implication of landscape changes using google earth historical imagery in Yenagoa Bayelsa State, Nigeria. Journal of Scientific Research, 2022, 5(1), 20-31. DOI: 10.47752/sjsr.51.20.31

[15] Bamiekumo BP, Akpobome EO, Kemebaradikumo AN, Mene-Ejegi OO, Eteh DR. Machine learning-based flood extent mapping and damage assessment in Yenagoa, Bayelsa State, using Sentinel-1 and 2 imagery (2018-2022). Discovery Nature, 2025, 2(3), e2dn1041. DOI: 10.54905/disssi.v2i3.e2dn1041

[16] Eteh DR, Egobueze FE, Paaru M, Otutu A, Osondu I. The impact of dam management and rainfall patterns on flooding in the Niger Delta: using Sentinel-1 SAR data. Discover Water, 2024, 4, 123. DOI: 10.1007/s43832-024-00185-8

[17] Eteh DR, Japheth BR, Akajiaku CU, Osondu I, Mene-Ejegi OO, Nwachukwu EM, et al. Assessing the impact of climate change on flood patterns in downstream Nigeria using machine learning and geospatial techniques (2018–2024). Discover Geoscience, 2025, 3(1), 76. DOI: 10.1007/s44288-025-00178-7

[18] Jonathan LE, Winston AG, Chukwuemeka P. Machine Learning and Morphometric Analysis for Runoff Dynamics: Enhancing Flood Management and Catchment Prioritization in Bayelsa, Nigeria. Journal of Computational Systems and Applications, 2025, 2(2), 1-6. DOI: 10.63623/kkx1m906
[19] City Population. Yenagoa (Nigeria) population statistics. 2022. https://citypopulation.de/en/nigeria/
[20] Doust H, Omatsola E. Niger Delta. In: Edwards JD, Santogrossi PA Eds., Divergent/Passive Margin Basins, American Association of Petroleum Geologists, Tulsa, 1990, 239-248. DOI: 10.1306/M48508C4
[21] Short KC, Stäuble AJ. Outline of geology of Niger Delta. AAPG bulletin, 1967, 51(5), 761-779. DOI: 10.1306/5D25C0CF-16C1-11D7-8645000102C1865D
[22] Oborie E, Fatunmibi I, Otutu AO. Shoreline change assessment in the Orashi River, Rivers State, Nigeria, using the digital shoreline analysis system (DSAS). Sumerianz Journal of Scientific Research, 2023, 6(4), 70-77. DOI:10.47752/sjsr.64.70.77
[23] Okpobiri O, Akajiaku CU, Eteh DR, Moses P. Using machine learning and GIS to monitor sandbars along the River Niger in the Niger Delta, Nigeria. International Journal of Environment and Climate Change, 2025, 15(2), 182-203. DOI: 10.9734/ijecc/2025/v15i24721
[24] Yang S, Luo D, Tan J, Li S, Song X, Xiong R, et al. Spatial mapping and prediction of groundwater quality using ensemble learning models and shapley additive explanations with spatial uncertainty analysis. Water, 2024, 16(17), 2375. DOI: 10.3390/w16172375
[25] Wegahita NK, Ma L, Liu J, Huang T, Luo Q, Qian J. Spatial assessment of groundwater quality and health risk of nitrogen pollution for shallow groundwater aquifer around Fuyang city, China. Water, 2020, 12(12), 3341. DOI: 10.3390/w12123341
[26] merican Public Health Association. Standard methods for the examination of water and wastewater.23rd ed., American Public Health Association, American Water Works Association and Water Environment Federation, 2017, pp. 1976.
[27] Kressy DG. Prediction of Abidjan groundwater quality using machine learning approaches: An exploratory study. Intelligent Control and Automation, 2024, 15(4), 215-248. DOI: 10.4236/ica.2024.154010
[28] Wang J, Yan H, Xin K, Tao T. Risk assessment methodology for iron stability under water quality factors based on fuzzy comprehensive evaluation. Environmental Sciences Europe, 2020, 32(1), 81. DOI: 10.1186/s12302-020-00356-z
[29] Karimi H, Sahour S, Khanbeyki M, Gholami V, Sahour H, Shahabi-Ghahfarokhi S, et al. Enhancing groundwater quality prediction through ensemble machine learning techniques. Environmental Monitoring and Assessment, 2024, 197(1), 21. DOI: 10.1007/s10661-024-13506-0
[30] Nourani V, Ghaffari A, Behfar N, Foroumandi E, Zeinali A, Ke CQ, et al. Spatiotemporal assessment of groundwater quality and quantity using geostatistical and ensemble artificial intelligence tools. Journal of Environmental Management, 2024, 355, 120495. DOI: 10.1016/j.jenvman.2024.120495
[31] Dritsas E, Trigka M. Efficient data-driven machine learning models for water quality prediction. Computation, 2023, 11(2), 16. DOI: 10.3390/computation11020016
[32] Ahmad T, Aziz MN. Data preprocessing and feature selection for machine learning intrusion detection systems. ICIC Express Letters, 2019, 13(2), 93-101. DOI: 10.24507/icicel.13.02.93
[33] Jonathan EL, Imoni O, Chukwuemeka P, Eteh DR. Impact of oil spills on mangrove ecosystem degradation in the Niger Delta using remote sensing and machine learning. Journal of Geography and Cartography, 2025, 8(2), 11707. DOI: 10.24294/jgc11707
[34] Lantz B. Machine Learning with R (4th Edition) - Learn Techniques for Building and Improving Machine Learning Models, from Data Preparation to Model Tuning, Evaluation, and Working with Big Data. Packt Publishing, 2023. https://app.knovel.com/kn/resources/kpMLRL0001/toc
[35] Awad M, Khanna R. Efficient learning machines: theories, concepts, and applications for engineers and system designers. Springer Nature; 2015. DOI: 10.1007/978-1-4302-5990-9
[36] Pereira GW, Valente DS, Queiroz DM, Coelho AL, Costa MM, Grift T. Smart-map: an open-source QGIS plugin for digital mapping using machine learning techniques and ordinary kriging. Agronomy, 2022, 12(6), 1350. DOI: 10.3390/agronomy12061350
[37] Berhanu KG, Hatiye SD, Lohani TK. Coupling support vector machine and the irrigation water quality index to assess groundwater quality suitability for irrigation practices in the Tana sub-basin, Ethiopia. Water Practice and Technology, 2023, 18 (4), 884-900. DOI: 10.2166/wpt.2023.055
[38] Singh PK, Rajput J, Kumar D, Gaddikeri V, Elbeltagi A. Combination of discretization regression with data-driven algorithms for modeling irrigation water quality indices. Ecological Informatics, 2023, 75, 102093. DOI: 10.1016/j.ecoinf.2023.102093
[39] Rammohan B, Partheeban P, Ranganathan R, Balaraman S. Groundwater quality prediction and analysis using machine learning models and geospatial technology. Sustainability, 2024, 16(22), 9848. DOI: 10.3390/su16229848
[40] Apogba JN, Anornu GK, Koon AB, Dekongmen BW, Sunkari ED, Fynn OF, et al. Application of machine learning techniques to predict groundwater quality in the Nabogo Basin, Northern Ghana. Heliyon, 2024, 10(7), e28527. DOI: 10.1016/j.heliyon.2024.e28527
[41] Igwebuike N, Ajayi M, Okolie C, Kanyerere T, Halihan T. Application of machine learning and deep learning for predicting groundwater levels in the West Coast Aquifer System, South Africa. Earth Science Informatics, 2025, 18(1), 6. DOI: 10.1007/s12145-024-01623-w
[42] Mosavi A, Ozturk P, Chau KW. Flood prediction using machine learning models: Literature review. Water, 2018, 10(11), 1536. DOI: 10.3390/w10111536
[43] Chukwuemeka P, Kyrian O, Imoni O. Leveraging machine learning for the identification of Obfuscated javascript in phishing attacks. Asian Journal of Research in Computer Science, 2025, 18(6), 301-314. DOI: 10.9734/ajrcos/2025/v18i6700
[44] Rowland ED, Oseji S, Iziegbe E, Abaye ON, Oreikio E. Water quality assessment using GIS based multi-criteria evaluation (MCE) and analytical hierarchy process (AHP) methods in yenagoa bayelsa state, Nigeria. International Journal of Advanced Engineering Research and Science, 2023, 10(4). DOI: 10.22161/ijaers.104.9
[45] Berhanu KG, Lohani TK, Hatiye SD. Spatial and seasonal groundwater quality assessment for drinking suitability using index and machine learning approach. Heliyon, 2024, 10(9), e30362. DOI: 10.1016/j.heliyon.2024.e30362
[46] Shams MY, Elshewey AM, El-Kenawy ES, Ibrahim A, Talaat FM, Tarek Z. Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications, 2024, 83(12), 35307-35334. DOI: 10.1007/s11042-023-16737-4

[47] Akajiaku UC, Ohimain EI, Olodiama EE, Eteh DR, Winston AG, Chukwuemeka P, et al. Identifying suitable dam sites using geospatial data and machine learning: a case study of the katsina-ala river in Benue State, Nigeria. Earth Science Informatics, 2025, 18(3), 497. DOI: 10.1007/s12145-025-01974-y
[48] Gupta AN, Kumar D, Singh A. Evaluation of water quality based on a machine learning algorithm and water quality index for mid gangetic region (south Bihar plain), India. Journal of the Geological Society of India, 2021, 97(9), 1063-1072. DOI: 10.1007/s12594-021-1821-0
[49] Oseji S, Chukwuemeka P, Imoni O. Artificial intelligence in 3D printed concrete: Sustainability assessment and implementation challenges. Journal of Materials Science Research and Reviews, 2025, 8(2), 515-528. DOI: 10.9734/jmsrr/2025/v8i2421
[50] Matsui K, Kageyama Y. Water pollution evaluation through fuzzy c-means clustering and neural networks using ALOS AVNIR-2 data and water depth of Lake Hosenko, Japan. Ecological Informatics, 2022, 70, 101761. DOI: 10.1016/j.ecoinf.2022.101761
[51] Malakar P, Mukherjee A, Bhanja SN, Saha D, Ray RK, Sarkar S, et al. Importance of spatial and depth-dependent drivers in groundwater level modeling through machine learning. Hydrology and Earth System Sciences Discussions, 2020, 1-22. DOI: 10.5194/hess-2020-208
[52] Asmoay AA, Shams EM, Galal WF, Mohamed A, Sawires R. Geochemical characterization and health risk assessment of groundwater in Wadi Ranyah, Saudi Arabia, using statistical and GIS-based models. Environmental Geochemistry and Health, 2025, 47(6), 208. DOI: 10.1007/s10653-025-02517-6
[53] Zenebe GB, Hailu G, Girmay A, Hussien A, Abrehe S. Evaluation of geostatistical interpolation methods on spatial representation of groundwater depth and nitrate concentration of Elalla-Aynalem wellfield, Northern Ethiopia. Discover Water, 2025, 5(1), 9. DOI: 10.1007/s43832-025-00189-y