Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal
AIMS Research Journal

Volume 10, No 4, December 2024 Series

Smart Water Quality Monitoring System's Performance

[1]Ihama P.O. & [2]Obahiagbon, K.O.
[1,2] Department of Computer Science
Benson Idahosa University
Edo State, Nigeria
Emails: osayandepeace.op@gmail.com; kobahiagbon@biu.edu.ng

ABSTRACT

Availability of clean and safe drinking water is one of the key determinants for public health and sustainable development. This study presents smart water quality monitoring (SWQM) system that applies sophisticated machine learning techniques for strengthening the accuracy of water quality assessments. IoT and Machine Learning Simulation Models, such as Random Forest and Gradient Boosting, have been used to determine the water portability when certain parameters such as PH, turbidity, and Dissolved Oxygen are kept as inputs into the SWQM system. Experimental studies, guided by methodologies such as Box-Behnken Design and Central Composite Design have made optimizations in coagulation processes used for improving urban drinking water treatment by manipulating the reduction of Total Organic Carbide, Total Nitrogen, and Total Suspended Solids concentrations. Real-time data collection and analysis efficiency using the enhanced IoT-enabled SWQMS is however proven to be superior and more effective with Random Forest model precision and recall. This study demonstrates the significance of taking IoT and ML into account when thinking about managing water resources. This research is necessary to solve the ground realities of developing countries concerning pollution and quality measurements in water monitoring.

Keywords: Smart Water Quality Monitoring (SWQM), Dissolve Oxygen, Support Vector Machine, Box – Behnken Design (BBD), Central Composite Design (CCD)

Aims Research Journal Reference Format:
Ihama P.O. & Obahiagbon, K.O. (2024): Smart Water Quality Monitoring System's Performance. Advances in Multidisciplinary and Scientific Research Journal Vol. 10. No. 4. Pp 31-52.. www.isteams.net/aimsjournal. dx.doi.org/10.22624/AIMS/V10N4P4

1. INTRODUCTION

Nigeria has experienced a prolonged history of water pollution challenges since its establishment in 1960. (Sabari *et al*., 2020). About 66.3 million Nigerians do not have access to safe drinking water (Berry *et al*., 2019). Besides the pollution of the water at the sources, there is also a significant deterioration of its quality by the time it gets to the point of use due to improper handling (Berry *et al*., 2019). Over the years, the water environment in developing countries like Nigeria has suffered from pollution, with little attention paid to the environmental risks posed by unregulated growth on water quality. Despite the proactive efforts of the Nigerian government to manage water resources, the issue of water pollution remains a persistent concern. The study aimed to evaluate the effectiveness of the coagulation water treatment process in removing pollutants such as Total Organic Carbon (TOC), Total Nitrogen (TN), and Total Suspended Solids (TSS) from urban drinking water.

Polyaluminium Chloride (PAC) was used as a coagulant to assess the impact of the treatment process on the composition and diversity of these contaminants in metropolitan water supplies (Chen *et al*., 2022; Yateh *et al*., 2023). In the literature, an experimental design technique known as the Box-Behnken Design (BBD), was utilized to optimize multiple responses by varying three factors: pH, temperature (°C), and dosage (mgL^(-1)), each at three levels (low, medium, and high). A second-order quadratic regression model was employed to fit the water quality data, allowing for the capture of quadratic trends and the identification of optimal conditions for PAC performance.

2. SMART WATER QUALITY MONITORING (SWQM)

The SWQM comprises three components that collectively establish a fundamental network for the remote monitoring of water quality. These components encompass the sensing system, the communication system, and the head-end system (Yaroshenko et al., 2020). The Wireless Sensor Network sensing system executes the tasks of data collection, processing, and transmission. The process of data collection is facilitated through an array of sensing devices positioned at various locations within water bodies. This setup enables the collection of water samples over extensive areas at consistent time intervals (Yaroshenko *et al*., 2020).

The sensing module includes a sensor transducer that measures the parameter and sends it to the processing unit for further analysis; thereafter, the data is transmitted through a communication unit to the intermediate nodes or gateway (Yaroshenko et al., 2020). All these operations are enabled by the power supply unit. Deploying multiple sensors at various locations along water bodies to acquire samples at more frequent intervals enhances the precision of water quality assessments (Yaroshenko et al., 2020). The enhancement is attributed to the increased availability of data for water quality studies.

The communication system is responsible for transmitting the detected data to the head-end system. In a star architecture, this sensing node can transmit data directly to the gateway node, through intermediary nodes to the gateway node, or occasionally to the cloud. The gateway node facilitates simpler data transmission across a base station. The network topology, whether mesh or star, is the sole factor that affects the choice among different communication scenarios (Yaroshenko et al., 2020). Various network communication structures are available, categorized into three types: short-range, medium-range, and long-range communication. The enhancement is attributed to the increased availability of data for water quality studies.

The communication system is responsible for transmitting the detected data to the head-end system. In a star architecture, the sensing node can transmit data directly to the gateway node, indirectly through intermediary nodes to the gateway node, or occasionally to the cloud. The gateway node simplifies data transmission through a base station. The network topology, whether mesh or star, is the sole factor that impacts the choice between different communication scenarios (Yaroshenko et al., 2020). There are various network communication structures that can be classified into three categories: short-range, medium-range, and long-range communication. Moreover, the HES includes a user interface that performs additional computations, such as data classification and organization derived from the WSN.

Several methods are available for storing the acquired data, including offline, online, or cloud solutions. Data can be displayed to users using tables, charts, or graphs. Furthermore, supplementary calculations can be performed to visually depict water quality in water bodies by creating maps that illustrate the geographical distribution of water quality. Typically, remote monitoring stations archive water quality data in databases supported by management systems, which are mainly available online (Yaroshenko et al., 2020).

2.1 Summary of Reviewed Literature

Table 2.1: Summary of Reviewed Literature

S/ N	Author	Title	Year	Contributions to study	Limitations
1	Wiryasputra *et al.,*	IoT real-time potable water quality monitoring and prediction model grounded on cloud computing architecture to mitigate health risks.	2024	By leveraging IoT, particularly for water quality monitoring, data on water quality components like temperature, alkalinity/acidity, and contaminants were gathered using a network of sensors. Through the amalgamation of machine learning techniques and water quality data, real-time predictions concerning the current potable water quality status were made.	Enhanced sensors were not integrated or compared with alternative prediction models to fortify the overall monitoring system.

2	Murti *et al.,*	An intelligent system for monitoring water quality by leveraging long-range Internet of Things technology.	2024	Their study suggested a tool enabled by IoT technology to swiftly and effectively monitor water quality through pH and turbidity sensors. These sensors were interfaced with a microcontroller as a controller, linked to a cloud service named Antares for data storage, and displayed on an android platform.	The study did not incorporate the utilization of the water discharge parameter.
3	Lal *et al.,*	Cost-effective IoT-based system for monitoring lake water quality.	2024	Developed and tested a system equipped with low-cost sensors to measure fundamental water quality parameters such as turbidity, total dissolved solids, temperature, pH, and dissolved oxygen. The system integrated IoT technology, solar power, and the capability to float akin to a small boat in fresh water.	The system lacked a detailed system architecture

4	Yateh *et al*.	Application of Response Surface Methodology to Optimize Coagulation Treatment Process of Urban Drinking Water Using Polyaluminium Chloride	2023	Investigated the efficiency of the coagulation water treatment process to remove pollutants such as total organic carbon (TOC), total nitrogen (TN), and total suspended solids (TSS) from urban drinking water. The polyaluminium chloride (PAC) coagulant was applied to determine the impact of the treatment process on the structure and diversity of these pollutants in urban drinking water. Furthermore, the response surface methodology by the Box–Behnken optimization analysis was applied to coagulant dosage, temperature, pH using the Quadratic model.	The second-order quadratic model is limited at the boundaries of the fitted curve and the BBD lack star point that addresses local variability and rotation in the data.

5	Goodarzi *et al.,*	The estimation of water quality index using machine learning algorithms in a specific case study conducted in the Yazd-Ardakan Plain, Iran.	2023	Utilized WQI (WHO) and Fuzzy AHP-WQI methods to assess the quality of 96 wells in the area, and subsequently compared the outcomes of these two approaches. Results from the WQI (WHO) method revealed that 72 out of 96 wells were classified as having good water quality, while 23 wells were rated as poor.	The study did not delve into investigating uncertainties in critical values or weights within the WQI metric, and lacked a detailed system architecture.
6	Alzahrani *et al.,*	Internet of things (IoT)-based wastewater management in smart cities	2023	The simulated analysis demonstrated that the proposed approach attains a high wastewater recycling rate of 96.3%, efficiency ratio of 88.7%, low moisture content ratio of 32.4%, increased wastewater reuse of 90.8%, and prediction ratio of 92.5%.	Deep learning technology was not incorporated for expanding the system.

7	Shams *et al.,*	The utilization of machine learning models based on the grid search method for water quality prediction.	2023	The grid search approach was employed to tune parameters for four classification models and four regression models. RF, XGBoost, AdaBoost, and GB models were used for classification, while KNN, DT, SVR, and MLP models were used for regression in predicting WQC and WQI, respectively. Assessment metrics such as accuracy, recall, precision, F1 score, MCC, MAE, MedAE, MSE, and R2 were computed to evaluate model performance.	Recurrent neural networks with LSTM were not employed in the prediction, nor was a time series analysis of WQI and WQC in the presence of climate change variables conducted.
7	Jáquez *et al.,*	An expansion of LoRa coverage and the incorporation of an unsupervised anomaly detection algorithm into an Internet of Things (IoT) system designed for monitoring water quality.	2023	The system undertook tasks such as data collection, storage, anomaly detection, and remote real-time alarm transmission to enable information.	The proposed system lacked a detailed system architecture.

8	Chen *et al.,*	An IoT-Based Fish Farm Water Quality Monitoring System.	2022	A robotic arm was engineered for executing automatic measurements and maintenance tasks, featuring a programmable logic controller, a single chip integrated with a wireless transmission module, and an embedded system. The system was segregated into control, measurement, server, and mobility components.	The authors did not employ the grouper model farm and big data for integrating diverse monitoring modules in breeding ponds.
9	Singh *et al.,*	A study on water quality monitoring and management within building water tanks through the utilization of Industrial Internet of Things (IoT) technologies.	2021	Their proposal centered on an IoT-enabled framework for monitoring both water level and quality in domestic water tanks, featuring distinct upper and lower tank monitoring units. Integration of a cloud server-enabled Virtuino app facilitated real-time monitoring and visualization of sensor data on a graphical user interface (GUI)	Absence of sensors like dissolved oxygen, conductivity, as well as the lack of an edge computing-enabled vision device for rapid detection of specific bacteria and harmful particles through machine learning algorithms.

10	Bogdan *et al.,*	A cost-effective Internet of Things (IoT) water-quality monitoring system tailored for rural regions.	2023	Findings from their study indicated the scalability of the system to cater to the water monitoring needs of different rural areas. Moreover, their experiments identified suitable water sources for public consumption while flagging those that should be avoided. Notably, all the tested water sources were potable, with an exception where total dissolved solids (TDS) exceeded the acceptable limit of 500 ppm.	The authors overlooked the addition of extra sensors, and a data analysis methodology grounded in various machine-learning techniques.

3. DESCRIPTION OF THE DATASET

Access to clean and safe drinking water is not only a fundamental human right but also a crucial determinant of public health and sustainable development. Recognizing the strong correlation between water quality and human well-being, this study investigates a dataset encompassing water quality metrics from 3276 diverse water bodies, available in kaggle, an online data repository. The dataset features a range of parameters indicative of water portability, including pH, hardness, total dissolved solids (TDS), chloramines, sulphate, conductivity, organic carbon, and turbidity. Each parameter's role in determining water safety is briefly outlined, referencing standards established by organizations like the World Health Organization (WHO) and the US Environmental Protection Agency (EPA). For instance, acceptable pH levels are noted as falling between 6.5 and 8.5, while TDS should ideally be below 500 mg/L.

The dataset further includes a binary 'Portability' indicator, classifying each water body as either safe (1) or unsafe (0) for human consumption. This categorization serves as the target variable for subsequent analyses, potentially enabling the development of predictive models to assess portability based on measured water quality metrics.

3.1 Sequence Diagram of the Current System

Figure 3.1 illustrates the sequence of steps and interactions involved in a real-time water quality monitoring system, likely using simulated data for demonstration or testing purposes.

Key Components:

i. Local Control Unit: This is the on-site control center where the monitoring process is initiated and potentially where local data visualization or alarms might be displayed.
ii. Remote Management Center: A central location where data is further analyzed and potentially where more comprehensive visualization and management tools are available.
iii. Central Control Module: The core module responsible for coordinating the entire monitoring process.
iv. Data Processing & Transmission Module: Handles the preprocessing of raw data and its transmission to other modules or centers.
v. Data Acquisition Module: Gathers data from various sensors (Conductivity, pH, Temperature, etc.).
vi. Wireless Module: Facilitates wireless communication between the local control unit and the remote management center.
vii. Sensors: Devices measuring specific water quality parameters.
viii. Water Distribution Unit: The physical system where water is being distributed and monitored.

Sequence of Events:

1. User Initiates Monitoring: The process starts when a user (likely at the Local Control Unit) triggers the monitoring session.
2. Data Stream Initiation: The system starts receiving simulated sensor data. This could involve reading from a pre-recorded dataset or generating data based on predefined patterns.
3. Data Acquisition: The Data Acquisition Module collects the incoming sensor data.
4. Data Processing & Transmission: The raw data is preprocessed (cleaned, normalized, etc.) and then sent to the Central Control Module and potentially also transmitted to the Remote Management Center via the Wireless Module.
5. Prediction and Analysis: The Central Control Module uses machine learning models to analyze the data and predict water quality metrics.
6. Alert and Visualization: Based on the analysis, the system may trigger alarms if any parameters exceed predefined thresholds. Additionally, the processed data and predictions are visualized on dashboards at both the Local Control Unit and the Remote Management Center.

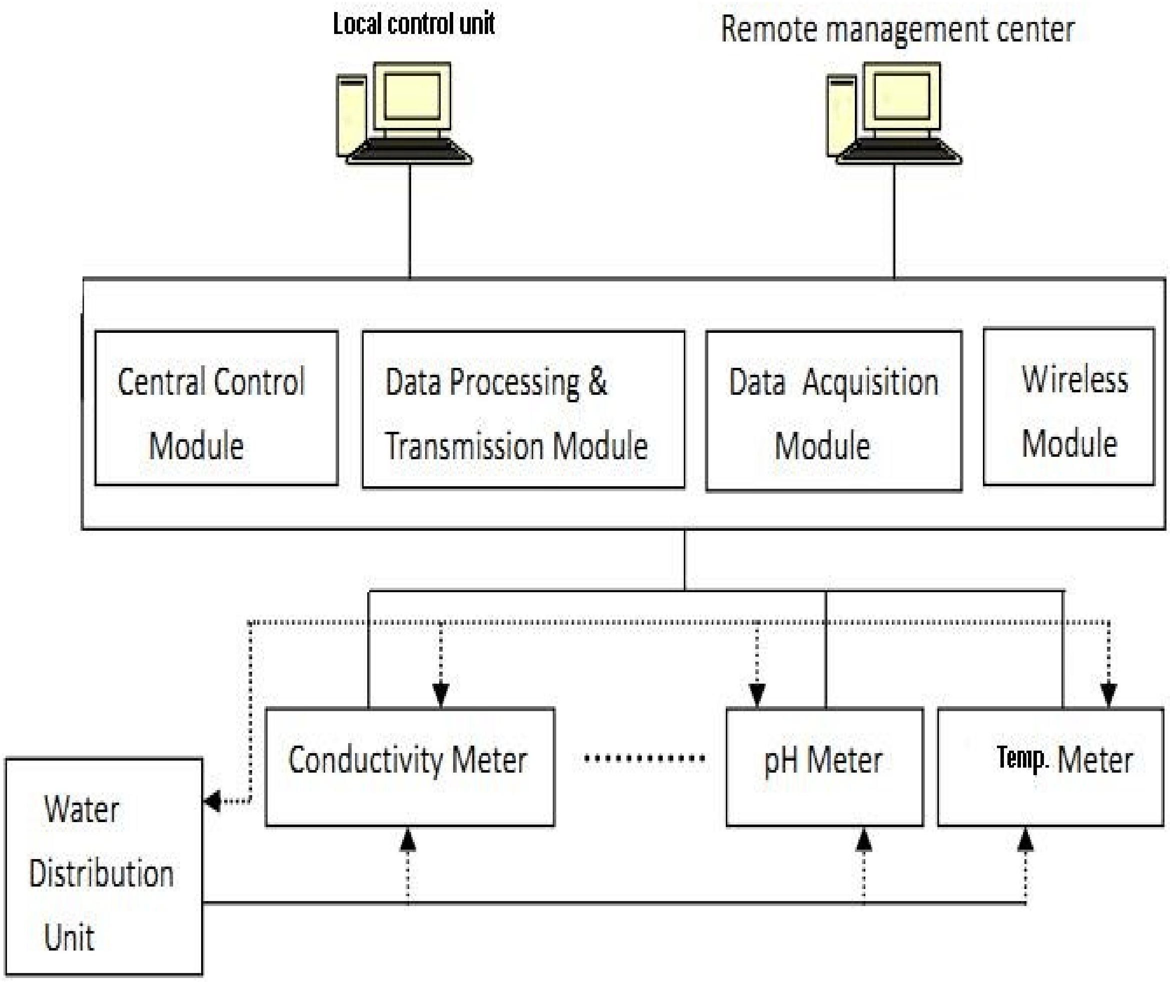

Figure 3.1: Sequence Diagram of the Current System

3.1 Class Diagram of the Current System

The class diagram in figure 3.2 illustrates the static structure of the system, illustrating the classes, characteristics, and methods involved:

i. Data Simulator: Handles the simulation of real-time water quality data.
ii. Data Processor: Responsible for cleaning, standardizing, and preparing data for analysis.
iii. Machine Learning Model: Implements the machine learning techniques used for predicting water quality.
iv. Alert System: Manages the development and dispatch of alerts based on the evaluation of water parameters.
v. Visualization Dashboard: Provides the interface for users to examine real-time data and analytics.

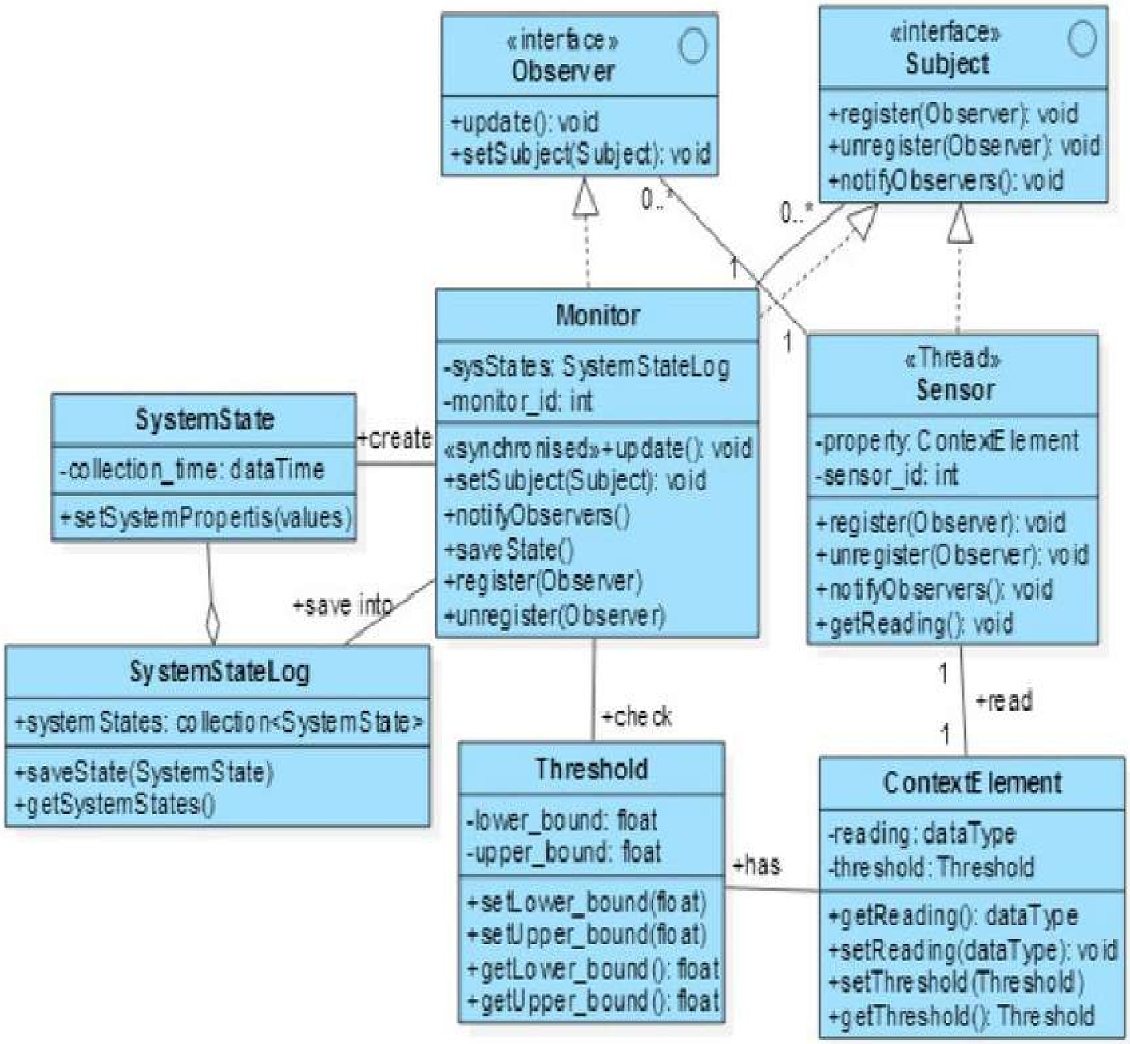

Figure 3.2: The Current System's Class Diagram

3.2 Experimental Design

In RSM application, the factors are usually more than one. Hence, the choice of appropriate levels to be studied for the explanatory variables is also vital as it can affect model correctness. The Experimental Design phase permits an appropriate design that can adequately and substantially estimation relationship between the response and one or more factors. Ordinarily applied DOEs in RSM include: 2^k full factorial design, 3^k full factorial design, and the CCD. In CCD, the number of experimental setup or run can be obtained by the mathematical relation given by; $2^k + 2k + k_c$, and all the factors are studied at five levels given as: $(-, \alpha, -1, 0, 1, \alpha)$, where 2^k is the full factorial design, $2k$ axial (star) points which are located at distance $\alpha = \sqrt[4]{2^k}$ from the center point and k_c. In this case $k = 3$, the numbers of factors utilized in the design and $k_c = 1$. Therefore, the total number of experimental run is equals thirteen and for the data collection see (Eguasa *et al*., 2022).

3.4.1 The Box – Behnken design (BBD)

A BBD permits for the design of the second-order regression model in a given response that is frequently used for process optimization (Hovat *et al*., 2013). The BBD comprises three types of trials namely; two levels (2^k) full factorial designs, $2k$ axial (star) points and C_p, p^{th} central points (Bezerra *et al.,* 2008). The mathematical expression for the BBD is given as:

$$BBD = 2k^2 - 2k + K_r \quad (1)$$

where $2k^2$ is the factorial portion, $2k$ is the axial or star points and $K_r$is at least pth central points utilized in the design. In this design $k = 3$ and $K_r = 3$ which from equation (1) sum up to 15 experimental run.

Table 3.1: Input process factors form BBD (Yateh *et al*., (2023))

Factors	Unit	Code	Levels	
			Low	High
pH	-	x_1	5	7
Temperature	℃	x_2	21	22
Dosage	mgL^{-1}	x_3	5	80

$$Number\ of\ runs\ for\ BBD = 2K^2 - 2K + K_r, where\ K = 3, r = 3$$

K= number of factors; r = number of independent generators; K_r = the replicate number of the central point.

Table 3.2 Experimental matrix for the factors and three responses (Yateh *et al*., (2023))

Runs	pH x_1	Temp. (^{0}C) x_2	Dosage x_3	TOC		TN		TSS	
				O	P	O	P	O	P
1	6	21	42.5	2.02	4.12	1.69	2.12	97.8	78.6
2	6	21	42.5	1.54	4.12	1.43	2.12	65.3	78.6
3	5	21	80	1.74	-2.15	1.45	2.03	122.2	129.7
4	7	21	80	1.24	1.41	1.12	1.17	190.3	181.3
5	6	20	5	21.36	19.8	1.41	1.72	77.6	73.5
6	6	22	80	5.54	7.1	2.28	1.98	110.6	114.7
7	6	21	42.5	8.81	4.12	3.23	2.12	72.8	78.6
8	6	22	5	5	2.84	3.19	3.52	60.0	62.6
9	7	21	5	2.77	6.66	2.75	2.17	120.8	113.3
10	6	20	80	2.63	4.79	3.21	2.87	131.9	129.3
11	5	22	42.5	2.97	5.3	2.53	2.25	114.5	103.0
12	5	20	42.5	3.75	5.49	1.48	1.23	71.5	66.6
13	5	21	5	3.53	3.35	1.48	1.43	80.7	89.7
14	7	20	42.5	18.39	16.06	1.44	1.72	141.9	153.4
15	7	22	42.5	3.33	1.6	1.37	1.62	86.5	91.4

Note: O means Observed Values, P means Predicted Values for the respective TOC, TN and TSS.

Table 3.3: Experimental design (BBD) for TOC, TN and TSS removal

Experimental Run	pH x_1	Temp. (°C) x_2	Dosage (mgL^{-1}) x_3	TOC y_1	TN y_2	TSS y_3
1	-1	-1	0	2.02	1.69	97.8
2	+1	-1	0	1.54	1.43	65.3
3	-1	+1	0	1.74	1.45	122.2
4	+1	+1	0	1.24	1.12	190.3
5	-1	0	-1	21.36	1.41	77.6
6	+1	0	-1	5.54	2.28	110.6
7	-1	0	+1	8.81	3.23	72.8
8	+1	0	+1	5	3.19	60.0
9	0	-1	-1	2.77	2.75	120.8
10	0	+1	-1	2.63	3.21	131.9
11	0	-1	+1	2.97	2.53	114.5
12	0	+1	+1	3.75	1.48	71.5
13	0	0	0	3.53	1.48	80.7
14	0	0	0	18.39	1.44	141.9
15	0	0	0	3.33	1.37	86.5

Data transformation using central composite design (CCD) to RSM Data

The values of the explanatory variables are coded between 0 and 1. The data collected via a CCD is transformed by a mathematical relation:

$$x_{NEW} = \frac{Min(x_{OLD}) - x_0}{(Min(x_{OLD}) - Max(x_{OLD}))} \tag{2}$$

where x_{NEW} is the transformed value, x_0 is the target value that needed to be transformed in the vector containing the old coded value, represented as x_{OLD}, $\mathrm{Min}(x_{OLD})$ and $Max(x_{OLD})$ are the minimum and maximum values in the vector x_{OLD} respectively, (Eguasa *et al*, 2022).

Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal
AIMS Research Journal
Volume 10, No 4, December 2024 Series

Table 3.4: Input process factors for with the addition of axial points (CCD)

Operating Factors	Symbol	Coded Factors	Coded Levels				
			$-\alpha = -1.682$	-1(Low)	0(Medium)	+1(High)	$+\alpha =$ +1.682
pH	-	x_1	4	5	6	7	8
Temperature	℃	x_2	20.5	21	21.5	22	22.5
Dosage	mgL^{-1}	x_3	0	5	42.5	80	85

Table 3.4, explains the choice of CCD in the addition of axial point to the coded factors that can capture curvature and maintain rotatability in the data $\alpha = \pm\sqrt[4]{2^k}$, where k= the number of factors used in the design. Therefore, $\alpha = \pm 1.682$ see Eguasa, (2020).

Table 3.5: Experimental design (CCD) for TOC, TN and TSS removal

Experimental Run	pH x_1	Temp. (℃) x_2	Dosage (mgL^{-1}) x_3	TOC $y_1 Observed$	TN $y_2 Observed$	TSS $y_3 Observed$
1	-1	-1	-1	2.02	1.69	97.8
2	1	-1	-1	1.54	1.43	65.3
3	-1	1	-1	1.74	1.45	122.2
4	1	1	-1	1.24	1.12	190.3
5	-1	-1	1	21.36	1.41	77.6
6	1	-1	1	5.54	2.28	110.6
7	-1	1	1	8.81	3.23	72.8
8	1	1	1	5	3.19	60.0
9	-1.682	0	0	2.77	2.75	120.8
10	1.682	0	0	2.63	3.21	131.9
11	0	-1.682	0	2.97	2.53	114.5
12	0	1.682	0	3.75	1.48	71.5
13	0	0	-1.682	3.53	1.48	80.7
14	0	0	1.682	18.39	1.44	141.9
15	0	0	0	3.33	1.37	86.5

Target points$x_0: -1, -1, -1,;$ $Min(x_{OLD}): -1.682, -1.682, -1.682,;$ $Max(x_{OLD}): 1.682, 1.682, 1.682$

$$x_{NEW} = \frac{Min(x_{OLD}) - x_0}{\left(Min(x_{OLD}) - Max(x_{OLD})\right)}$$

$$Explanatory\ variable\ x_1 : x_{11} = \frac{-1.682 - (-1)}{((-1.682) - (1.682))} = 0.2030$$
$$Explanatory\ variable\ x_2 : x_{12} = \frac{-1.682 - (-1)}{((-1.682) - (1.682))} = 0.2030$$
$$Explanatory\ variable\ x_3 : x_{13} = \frac{-1.682 - (-1)}{\left((-1.682) - (1.682)\right)} = 0.2030$$

Table 3.6: Experimental design for TOC, TN and TSS removal

Experimental Run	pH x_1	Temp. (°C) x_2	Dosage (mgL^{-1}) x_3	TOC $y_1 Observed$	TN $y_2 Observed$	TSS $y_3 Observed$
1	0.2030	0.2030	0.2030	2.02	1.69	97.8
2	0.7970	0.2030	0.2030	1.54	1.43	65.3
3	0.2030	0.7970	0.2030	1.74	1.45	122.2
4	0.7970	0.7970	0.2030	1.24	1.12	190.3
5	0.2030	0.2030	0.7970	21.36	1.41	77.6
6	0.7970	0.2030	0.7970	5.54	2.28	110.6
7	0.2030	0.7970	0.7970	8.81	3.23	72.8
8	0.7970	0.7970	0.7970	5	3.19	60.0
9	0.0000	0.5000	0.5000	2.77	2.75	120.8
10	1.0000	0.5000	0.5000	2.63	3.21	131.9
11	0.5000	0.0000	0.5000	2.97	2.53	114.5
12	0.5000	1.0000	0.5000	3.75	1.48	71.5
13	0.5000	0.5000	0.0000	3.53	1.48	80.7
14	0.5000	0.5000	1.0000	18.39	1.44	141.9
15	0.5000	0.5000	0.5000	3.33	1.37	86.5

Advances In Multidisciplinary & Scientific Research
AIMS Research Journal
A Multidisciplinary & Interdisciplinary Journal
Volume 10, No 4, December 2024 Series

Based on the type of response, the desirability function transforms the estimated response, $\hat{y}_p(\boldsymbol{x})$ to different individual scalar measure, $d_p\left(\hat{y}_p(\boldsymbol{x})\right)$ namely:

For larger-the-better (LTB) response $d_p\left(\hat{y}_p(\boldsymbol{x})\right)$ is given as:

$$d_p\left(\hat{y}_p(\boldsymbol{x})\right) \quad = \quad \begin{cases} 0, & \hat{y}_p(\boldsymbol{x}) < L \\ \left\{\frac{\hat{y}_p(x)-L}{T-L}\right\}^{t_1}, & L \leq \hat{y}_p(\boldsymbol{x}) \leq T, \\ 1, & \hat{y}_p(\boldsymbol{x}) > T, \end{cases} \qquad s.t\ \boldsymbol{x}\epsilon\ \varphi\ , \qquad (3)$$

where $T$and L are the maximum acceptable value and lower limit, respectively, of the p^{th} response. where ρ is the target value of the p^{th} response. However, for RSM data, the parameters values of $t_1\ and\ t_2$ are weights taken to be 1 for linearity (Eguasa *et al*., 2022).

4. OVERVIEW OF RESULTS

The water quality monitoring system's performance was evaluated using multiple machine learning models on a dataset that included various water portability parameters. The models tested included Logistic Regression, Random Forest, Support Vector Machine (SVC), K-Nearest Neighbors, Gradient Boosting, and a Neural Network (MLPClassifier). The evaluation metrics used were accuracy, precision, recall, F1-score, and ROC-AUC. Among the models, Random Forest and Gradient Boosting demonstrated superior performance, particularly in terms of accuracy and precision, which are crucial for ensuring reliable predictions of water portability.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.5250	0.5246	0.5325	0.5285	0.5439
Random Forest	0.7113	0.7302	0.6700	0.6988	0.7815
Support Vector Machine	0.6500	0.6493	0.6525	0.6509	0.7233
K-Nearest Neighbors	0.6425	0.6370	0.6625	0.6495	0.6814
Gradient Boosting	0.6475	0.6521	0.6325	0.6421	0.7091
Neural Network	0.6350	0.6292	0.6575	0.6430	0.7063
Voting Classifier	0.6900	0.6929	0.6825	0.6877	0.7648

Figure 4.1: First Code Evaluation

```
Results for Voting Classifier:
  Accuracy: 0.6900
  Precision: 0.6929
  Recall: 0.6825
  F1-Score: 0.6877
  ROC-AUC: 0.7648375000000001
```

Figure 4.2: Figure Showing Voting Classifier After Training

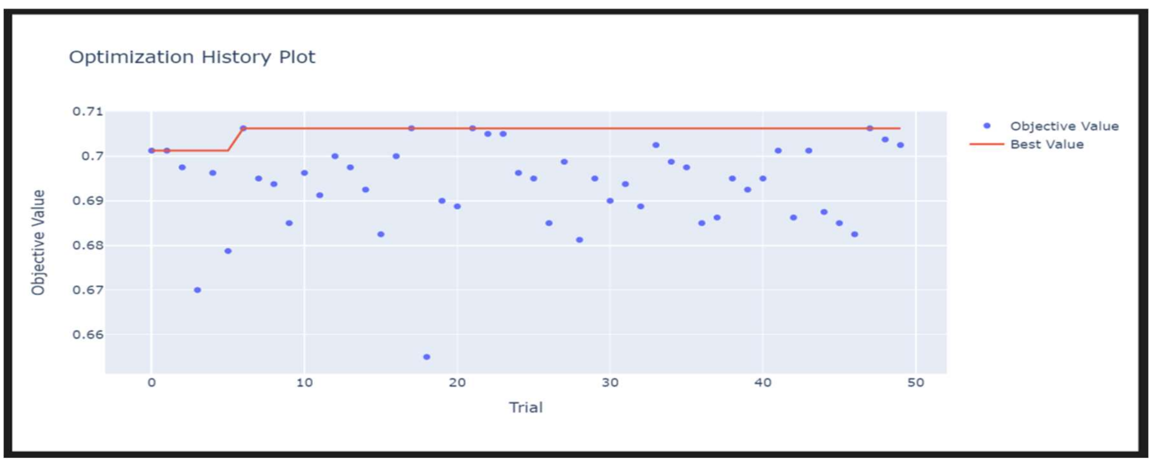

Figure 4.3: Optimization History Plot

Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal
Volume 10, No 4, December 2024 Series

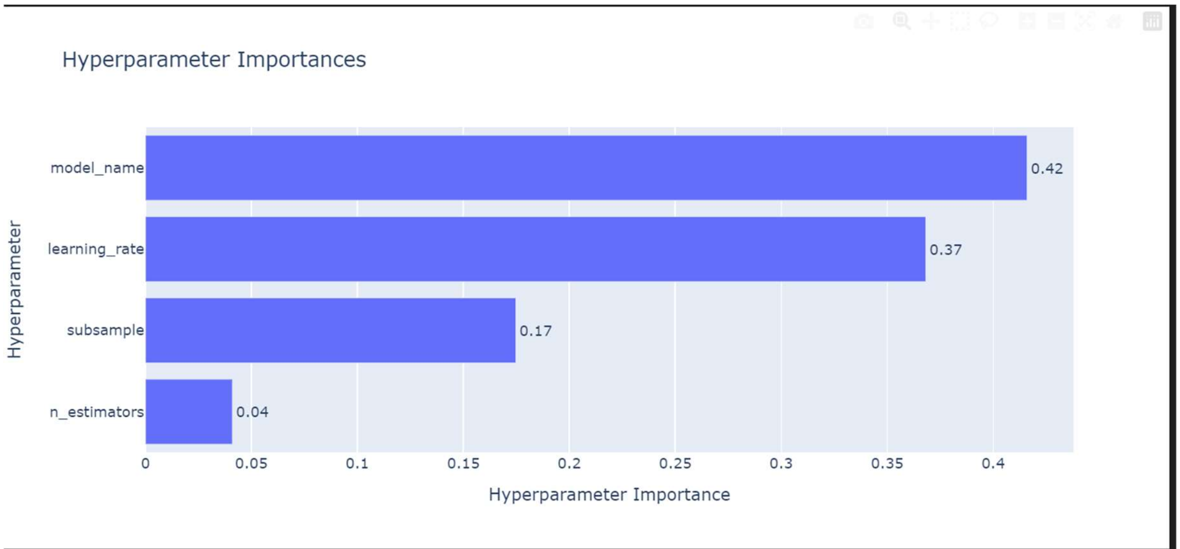

Figure 4.4: Hyperparameter Functions Importance

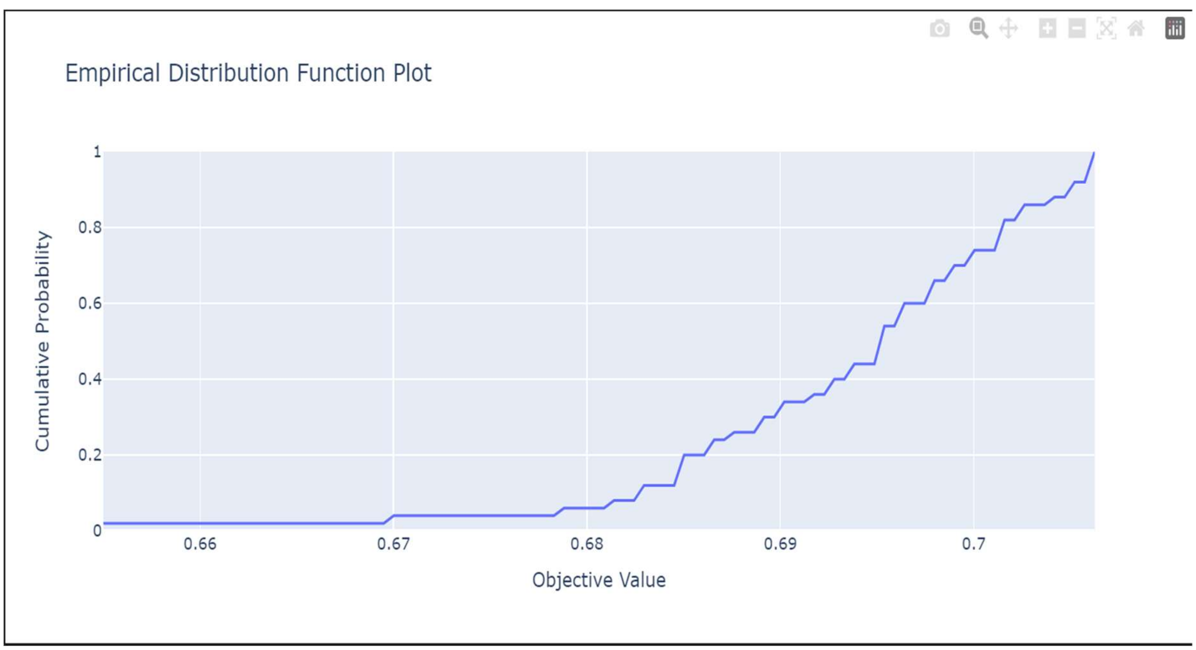

Figure 4.5:Final Empirical Distribution Training

4.31 Comparative Analysis of the various ML Models

A comparative analysis of the machine learning models revealed the following insights:

i. Random Forest emerged as one of the best-performing models, achieving high accuracy and balanced recall. The model's ability to aggregate the predictions of multiple decision trees made it robust against overfitting, which is often a challenge in classification tasks with diverse and potentially noisy data.
ii. Gradient Boosting also performed strongly, particularly after hyperparameter optimization using Optuna. The iterative nature of Gradient Boosting, where each tree attempts to correct the errors of its predecessor, allowed for fine-tuning that significantly improved predictive power. This model was particularly effective at identifying nuanced patterns in the data.
iii. Support Vector Machine (SVC) and K-Nearest Neighbors (KNN) provided strong precision scores, indicating their effectiveness in correctly identifying positive cases of potable water. However, they showed lower recall, suggesting a tendency to miss some instances, which could be critical in ensuring safe drinking water.
iv. Neural Network (MLPClassifier) showed competitive accuracy after considerable computational effort and hyperparameter tuning. Despite its complexity, the model's performance was comparable to the ensemble methods, making it a viable option in scenarios where the dataset is large and complex patterns need to be captured.

Overall, Random Forest and Gradient Boosting stood out as the most reliable models, offering a balance between sensitivity (recall) and specificity (precision), which is crucial for accurate water quality monitoring.

4.2 Sensor Performance

Although this study did not directly involve IoT sensors, the findings are highly relevant to real-world sensor applications in water quality monitoring. The robustness of the Random Forest and Gradient Boosting models suggests that these algorithms could effectively handle real-time data from IoT sensors, even in the presence of minor inaccuracies or noise. In a practical deployment, sensor calibration and maintenance would be critical to ensure data accuracy. The models' ability to manage noisy or imperfect data highlights their potential in real-time monitoring systems where sensor reliability might vary. Therefore, integrating these machine learning models with IoT-based monitoring systems could significantly enhance the accuracy and reliability of water quality assessments.

5. CONCLUSION

In conclusion, our work underlines the transformational potential of IoT and machine learning in water quality monitoring. The proposed simulation-based technique offers a practical and cost-effective way to solve the limitations of previous methodologies. The higher performance of Random Forest and Gradient Boosting models, along with their capacity to handle real-time sensor input, shows their applicability for real-world applications. The findings derived from the dataset analysis underscore the need of monitoring critical metrics including pH, turbidity, and dissolved oxygen for early detection of water quality issues. The incorporation of powerful machine learning algorithms into Internet of Things based monitoring systems holds the possibility of more effective water resource management and protection.

This research illustrates that the coagulation technique, recognized for its straightforwardness and cost-effectiveness, serves as a viable method for eliminating contaminants from urban drinking water. Although various coagulants have been evaluated for urban water treatment, Polyaluminium Chloride (PAC) emerged as particularly proficient in decreasing total organic carbon (TOC), total suspended solids (TSS), and total nitrogen (TN). As a result, the statistical modeling and optimization of the coagulation process were investigated in greater depth. The data gathered were analyzed using Response Surface Methodology (RSM) with a Central Composite Design (CCD) to ascertain optimal conditions and process specifications.

REFERENCES

Alzahrani, A. I. A., Chauhdary, S. H., and Alshdadi, A. A. (2023). Internet of Things (IoT)-Based wastewater management in smart cities. *Electronics*, *12*(12), 2590.

Berry, M. W., Mohamed, A.H., and Yap, B.W. (2019). Supervised and Unsupervised Learning for Data Science, Springer, Switzerland, 2019.

Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S., L. A. Escaleira, L. A., 2008. Response surface methodology (RSM) as a tool for optimization in analytical Chemistry. Talanta, 76, 965-977.

Bogdan, R., Paliuc, C., Crisan-Vida, M., Nimara, S., and Barmayoun, D. (2023). Low-Cost Internet-of-Things Water-Quality monitoring system for rural areas. *Sensors*, *23*(8), 3919.

Chen, C., Wu, Y., Zhang, J., and Chen, Y. (2022). IOT-Based Fish Farm Water Quality Monitoring System. *Sensors*, *22*(17), 6700.

Chen, C., Wu, Y., Zhang, J., and Chen, Y. (2022). IOT-Based Fish Farm Water Quality Monitoring System. *Sensors*, *22*(17), 6700.

Eguasa, O., Edionwe, E. and Mbegbu, J. I. (2022). Local Linear Regression and the problem of dimensionality: A remedial strategy via a new locally adaptive bandwidths selector, *Journal of Applied Statistics,* 50(6): 1283 – 1309.

Goodarzi, Mohammad Reza, Amir Reza R. Niknam, Ali Barzkar, Majid Niazkar, Yahia Zare Mehrjerdi, Mohammad Javad Abedi, and Mahnaz Heydari Pour. 2023. "Water Quality Index Estimations Using Machine Learning Algorithms: A Case Study of Yazd-Ardakan Plain, Iran" *Water* 15, no. 10: 1876. https://doi.org/10.3390/w15101876

Jáquez, A. D. B., Herrera, M. T. A., Celestino, A. E. M., Ramírez, E. N., and Cruz, D. a. M. (2023). Extension of LORA coverage and integration of an unsupervised anomaly detection algorithm in an IoT water quality monitoring system. *Water*, *15*(7), 1351.

Lal, K., Menon, S., Noble, F., and Arif, K. M. (2024). Low-cost IoT based system for lake water quality monitoring. *PloS One*, *19*(3), e0299089.

Mutri, M. A., Saputra, A. R. A., Alinursafa, I., Ahmed, A. N., Yafouz, A., and El-Shafie, A. (2024). Smart system for water quality monitoring utilizing long-range-based Internet of Things. *Applied Water Science*, *14*(4).

Sabari, M., Aswinth, P., Karthik, T., and Kumar C. (2020). Water Quality Monitoring System Based On IoT. 5th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India. 279-282.

Shams, M. Y., Elshewey, A. M., El-Kenawy, E. M., Ibrahim, A., Talaat, F. M., and Tarek, Z. (2023). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*.

Singh, R., Baz, M., Gehlot, A., Rashid, M., Khurana, M., Akram, S. V., Alshamrani, S. S., and AlGhamdi, A. S. (2021). Water quality monitoring and management of building water tank using industrial internet of things. *Sustainability*, *13*(15), 8452.

Wiryasaputra, R., Huang, C., Lin, Y., and Yang, C. (2024). An IoT Real-Time Potable water quality monitoring and prediction model based on cloud computing architecture. *Sensors*, *24*(4), 1180.

Yaroshenko, I., Kirsanov, D., Marjanovic, M., Lieberzeit, P. A., Korostynska, O., Mason, A. and Legin, A. (2020). Real-time water quality monitoring with chemical sensors. Sensors. 20(12). 3432.

Yateh, M., Lartey-Young, G., Li, F., Li., M. and Tang, Y. (2023). Application of Response Surface Methodology to Optimize Coagulation Treatment Process of Urban Drinking Water Using Polyaluminium Chloride. *Water*, *Vol. 15* (853), 1 – 13.

Yateh, M., Lartey-Young, G., Li, F., Li., M. and Tang, Y. (2023). Application of Response Surface Methodology to Optimize Coagulation Treatment Process of Urban Drinking Water Using Polyaluminium Chloride. *Water*, *Vol. 15* (853), 1 – 13.