Mesopotamian journal of Big Data
Vol. **(2024)**, 2024, **pp**. 211–222
DOI: https://doi.org/10.58496/MJBD/2024/015 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Automated Water Quality Assessment Using Big Data Analytics

Yasmin Makki Mohialden [1,*], Nadia Mahmood Hussien [1], Saba Abdulbaqi Salman [3]

[1] *Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq*

[2] *Department of Computer Science, College of Education, Al-Iraqia University, Baghdad, Iraq*

ARTICLE INFO

Article History

Received 15 Jul 2024
Accepted 13 Oct 2024
Published 07 Nov 2024

Keywords

Water Quality
Big Data
Environmental Contamination
Machine Learning
Random Forest
Gradient Boosting
Support Vector Machine

ABSTRACT

Water is one of the world's most precious resources, essential to life. Industrial waste, agricultural runoff, and urban discharge degrade water, rendering it unfit for consumption. Water quality monitoring and evaluation are more important than ever. Big Data analytics is used to examine water quality utilizing enormous datasets of pH, hardness, solids concentration, chloramine, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. This work classifies water potability, which is vital for human consumption, using strong machine learning on massive datasets. Classifiers were Random Forest, Gradient Boosting, and Support Vector Machine on 3,276 water bodies. The Random Forest classifier obtained the highest accuracy at 66.77% after significant data preparation and training, followed by Gradient Boosting at 66.01% and SVM at 62.80%. This shows that Big Data analytics and machine learning algorithms can interpret complex water quality data for public health and natural resource management.

The Random Forest classifier and SVM in this study accurately calculate water potability. Prediction algorithms consider water cleanliness data and may aid public safety and water resource monitoring.

1. INTRODUCTION

Water quality impacts human and environmental health since it's essential to life. Environmental contamination, industrial activities, and climate change complicate water quality monitoring. Traditional monitoring methods cannot handle these difficulties, necessitating more complicated, scalable solutions.

Massive and complex water quality monitoring datasets may be handled via Big Data analytics. Big Data solutions use powerful mathematical algorithms to manage massive volumes of irregularly organized data, providing deeper insights and more accurate projections. Traditional machine learning (ML) models are useful despite biases and overfitting. Advanced ML algorithms like bagging and boosting overcome these constraints, improving predictions[1,2,21].

This paper examines Big Data analytics with Random Forest, Gradient Boosting, and SVM. We evaluate boosting, an ensemble learning strategy that iteratively corrects model defects, using large datasets. This includes the supervised learning approach SVM, which can classify and regress complex datasets but performs better with fewer data samples.[3, 4,19]

Big Data analytics is used to assess water quality by monitoring pH, hardness, and chloramine concentrations. The main goal is to find the best water potability model to preserve human health and the environment. The project intends to improve freshwater resource management and public health decision-making by automating water cleanliness tests efficiently and reliably[22].

As environmental factors become more complicated, traditional water quality monitoring methods fail. These methods often lack the accuracy and scalability to manage enormous water quality datasets. Big Data analysis is needed for better water quality insights.

This work automates and improves water quality predictions using Big Data and machine learning. Through ML classifier analysis of massive datasets, the study shows how Big Data may improve public health and environmental decision-making. Because the Random Forest classifier predicts water potability better, this automated water quality monitoring

*Corresponding author. Email: ymmiraq2009@uomustansiriyah.edu.iq

system is dependable. These findings may enhance freshwater resource management and provide safe drinking water[20,27,28]].

The paper is structured as follows: Section 2 discusses related work, Section 3 details the proposed method, and Section 4 conclusions.

2. RELATED WORK

In 2023,a Machine learning model may be an effective water quality evaluation approach. The dataset, which includes TDS, BOD, pH, and dissolved oxygen, was used to evaluate the Random Forest model, which had the highest accuracy (78.96%). The SVM model was inaccurate. The study reveals that precision-recall curves may predict water quality and provide drinking water access[6].

To effectively measure water quality, another 2022 study examined AI methods using long-term WQ data. This study compared CATBoost, XGBoost, MLP, SVM, RF, LR, Decision Tree, and stacking ensemble models. ROC and precision-recall curves assessed classifier performance on 1,679 specimens over nine years. At 94.51%, CATBoost was the most accurate classifier. By layering all classifiers, the meta-classifier approached 100% accuracy.

The study suggests artificial intelligence use the CATBoost algorithm to classify water quality and improve drinking water quality predictions[7].

In 2021, research anticipated five WQIs over two geographical periods. The datasets were classified as "Very Poor" due to unequal water sample distribution, indicating data quality issues. This study classified water purity by pH, oxygen dissolution, conductivity, turbidity, stool coliform, and humidity using machine learning algorithms like Decision Tree (DT), k-nearest Neighbor (KNN), Logistic Regression (LogR), Multilayer Perceptron (MLP), and Naive Bayes. With 99% classification accuracy, the DT algorithm surpassed other models. Data collection restrictions including locations, parameter kinds, and water types cause bias and uncertainty water quality indexes, according to the study. The study shows that a more accurate index that incorporates topological, hydrological, and geographical temporal changes might enhance water purity assessment worldwide[8].

A 2020 research compared 10 learning models (7 conventional and 3 ensemble) for water quality prediction using 33,612 observations from China's main rivers and lakes from 2012 to 2018. The models were assessed for accuracy, recurrence, F1-score, and weighted F1-score. The study also examined water properties that may affect model projections. The results showed that larger datasets improve water quality prediction systems. Decision Tree (DT), Random Forest (RF), and Deep Cascade Forest (DCF) methods trained on pH, Dissolved Oxygen (DO), Chemical Oxygen Demand by permanganate (CODMn), and ammonia nitrogen (NH3–N) data predicted all six Chinese government water quality levels better than other models. In addition, DT, RF, and DCF found two important water indices (DO, CODMn, NH3–N; CODMn, NH3–N) with excellent specificity for water quality classification[9].

Two predicted scenarios were offered in 2019: one used upstream unit indicators to forecast results, while the other used water quality measures at each station. Both scenarios used the 12 parameters entered. Field data from 2009 and 2010 validated the WDT-ANFIS model, which beat all other models and improved predicting accuracy for all water quality indicators. Scenario 2 improved all water quality parameters by 0.5% to 5%. The model was validated and correctly predicted all water quality measures, with R^2 levels over 0.9[10].

3. PROPOSED METHOD

categorizes and predicts water quality based on several characteristics. A huge dataset collects conductivity, organic carbon content, trihalomethanes, pH, hardness, solids concentration, chloramines, and sulfate levels. We preprocess this large dataset before analysis to remove missing values, scale numerical properties, and encode categorical variables.

To train and evaluate models, data is split into training and testing sets using functions like `train_test_split` after preprocessing. We use advanced Big Data machine learning models due to the high data volume. We choose Random Forest, Gradient Boosting, SVM, and Voting Classifier to successfully analyze the large dataset[23-25].

These selected models discover Big Data's deep patterns and correlations on the training dataset. We measure the models' accuracy, precision, and recall using the testing dataset to evaluate their capacity to handle large-scale data post-training. We use a grid search or random search to tune hyperparameters to maximize model performance.

According to evaluation measures, the best model predicts real-life water quality. The Big Data technique involves retraining the model with new data to keep it accurate and effective. Feature significance analysis or SHAP values can reveal the model's judgments, enhancing transparency and interpretability. Finally, the strategy encourages experts and decision-makers to collaborate to improve water quality management by utilizing Big Data insights and ensuring safe drinking water. In this workflow, the Random Forest model predicted water quality best and processed and analyzed huge datasets most accurately[26]. Figure 1 presents the proposed system block diagram. Table 2 outlines the steps for

implementing the proposed method. Table 2 lists the algorithms used. Tables 3-5 showcase the data samples chosen for the experimental section, sourced from an extensive dataset of water parameters. Figure 2 shows the big data-driven machine learning workflow.

Figure 3 shows the proposed method for the Water Quality Analysis Use Case Diagram. For demonstration purposes, Figure 4 shows the distribution of water quality parameters. Figure 5 shows the correlation matrix of water quality parameters. Figure 6 illustrates the simulation of time-series data, specifically the water quality time series.

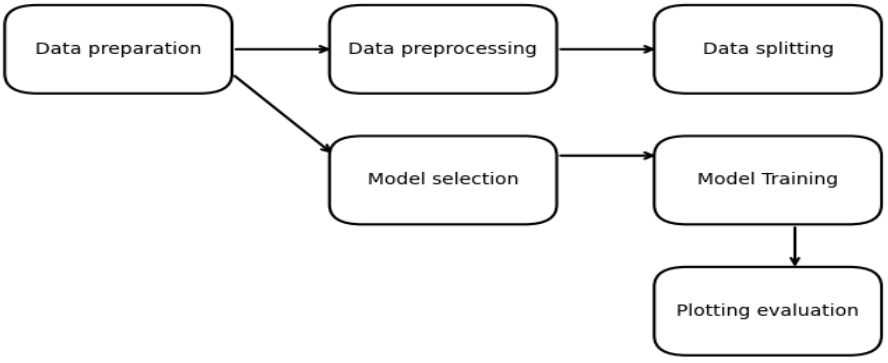

Fig .1. The block diagram of the proposed method.

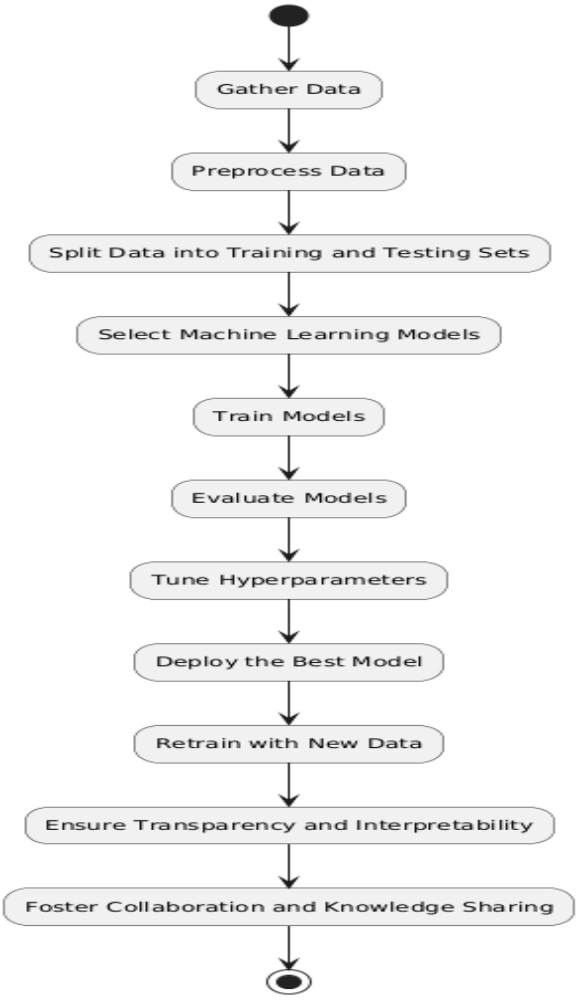

Fig .2. Big Data-Driven Machine Learning Workflow.

TABLE I. BIG DATA DRIVEN MACHINE LEARNING WORKFLOW FOR WATER QUALITY ANALYSIS AND PREDICTION

Step	Description	Code Implementation
1	Gathering Data: Establish a dataset with parameters related to water quality.	Define ideal parameter values and create a data frame with water quality data.
2	Preprocessing Data: Address missing values, scale numerical attributes, and encode categorical variables if needed.	Preprocess the data (not implemented in this snippet).
3	Splitting Data: Divide the dataset into training and testing sets.	Use train test split to split the data into training and testing sets.
4	Model Choice: Select appropriate machine learning models for classification tasks.	Choose models such as Random Forest, Gradient Boosting, SVM, and Voting Classifier.
5	Instruction of Models: Train machine learning models on the training dataset.	Fit the selected models on the training data.
6	Assessment of the Model: Evaluate the trained models using the testing dataset.	Calculate accuracy scores or other evaluation metrics using the testing data.
7	Tuning of Hyperparameters: Optimize model hyperparameters to improve performance.	Tune hyperparameters using methods like grid search or random search (not implemented in this snippet).
8	Model Deployment and Selection: Deploy the best-performing model for real-world predictions.	Select the best model based on evaluation metrics and deploy it for predictions.
9	Ongoing Enhancement: Retrain the model with updated data and continuously improve prediction accuracy.	Implement continuous improvement by retraining the model with new data periodically.
10	Transparency and Interpretability: Ensure model predictions are transparent and interpretable.	Provide explanations or insights into model decisions using feature importance analysis or SHAP values.
11	Cooperation and Exchange of Knowledge: Encourage collaboration and knowledge sharing among stakeholders.	Foster collaboration between experts, decision-makers, and communities to address water quality management.

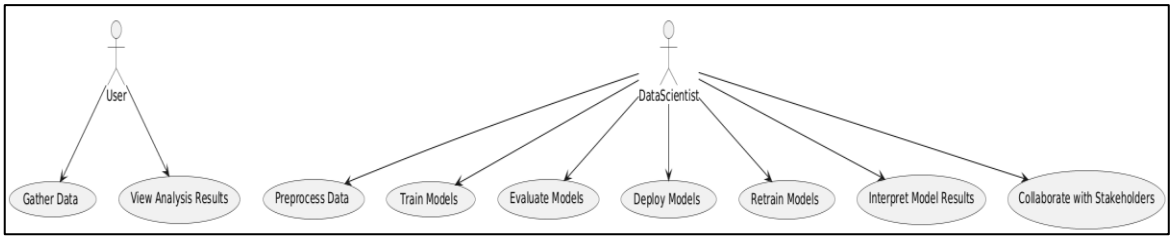

Fig .3. the proposed method for Water Quality Analysis Use Case Diagram.

TABLE II. MACHINE LEARNING ALGORITHMS USED FOR WATER QUALITY PREDICTION

Algorithm used	Description
Random Forest	Ensemble learning that builds several decision trees and outputs their mode (classification) or mean prediction (regression)[11,12].
Gradient Boosting	This ensemble learning approach produces a strong model by gradually adding weak learners (decision trees) and correcting previous model mistakes[13,14].
Support Vector Machine	A supervised learning method that builds hyperplanes in high-dimensional space to classify data into two groups by maximizing margin[15,16].
Voting Classifier	Ensemble learning uses numerous machine learning models to create a majority-voted prediction[17,18].

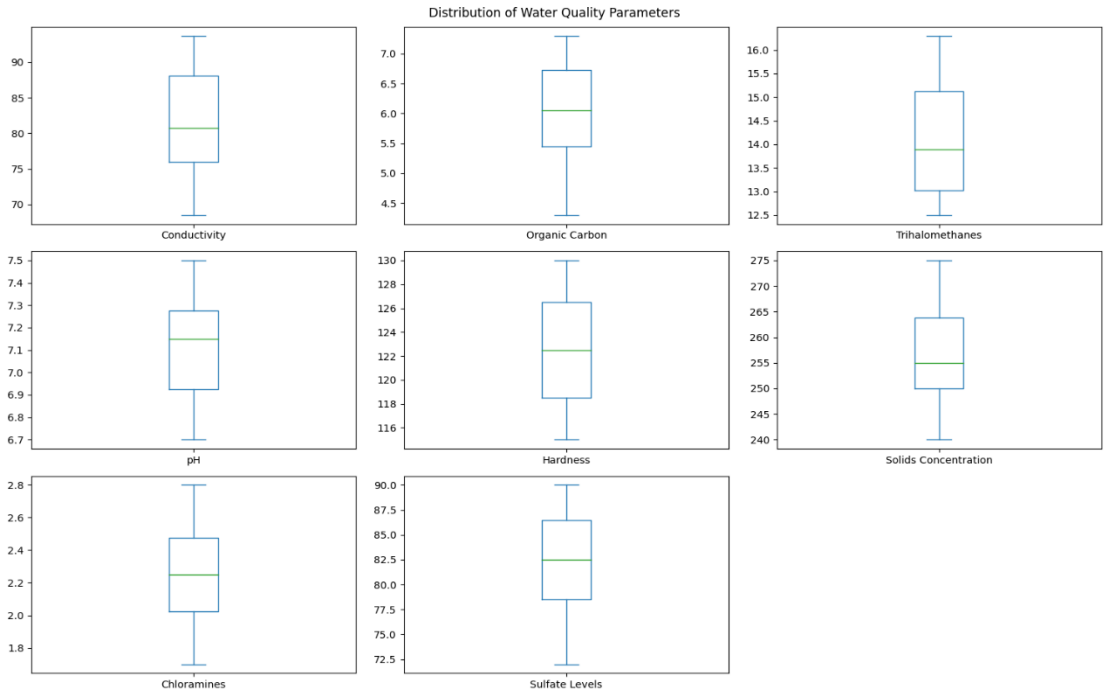

Fig .4. Distribution of Water Quality Parameters

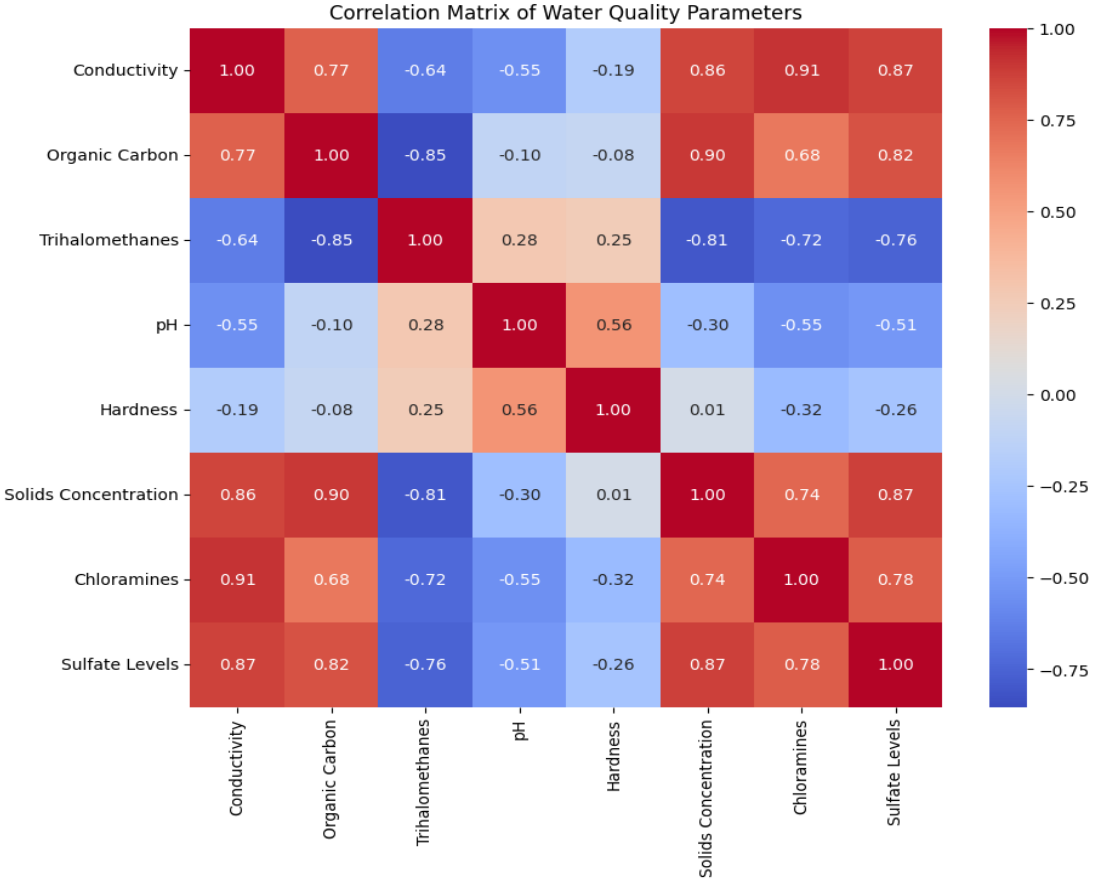

Fig .5. Correlation Matrix of Water Quality Parameters.

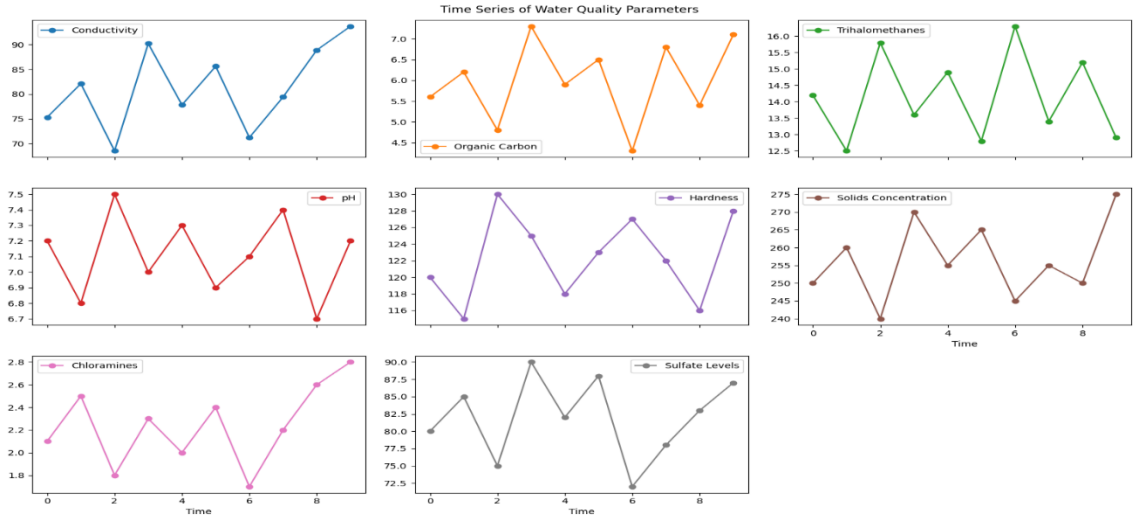

Fig .6. shows the water quality time-series simulation.

TABLE III. SAMPLE 1-WATER QUALITY PARAMETERS AND THEIR VALUES (SAMPLE SIZE OF 10)

Parameter	Description	Values
conductivity	Measure of water's ability to conduct electricity	[75.3, 82.1, 68.5, 90.2, 77.8, 85.6, 71.2, 79.4, 88.9, 93.7]
Organic carbon	Amount of carbon-containing compounds in water derived from living organisms	[5.6, 6.2, 4.8, 7.3, 5.9, 6.5, 4.3, 6.8, 5.4, 7.1]
trihalomethanes	Group of chemicals that may form when chlorine or other disinfectants are used	[14.2, 12.5, 15.8, 13.6, 14.9, 12.8, 16.3, 13.4, 15.2, 12.9]
pH	Measure of the acidity or basicity of a solution	[7.2, 6.8, 7.5, 7.0, 7.3, 6.9, 7.1, 7.4, 6.7, 7.2]
hardness	Measure of the concentration of calcium and magnesium ions in water	[120, 115, 130, 125, 118, 123, 127, 122, 116, 128]
Solids concentration	Total dissolved solids in water, including organic and inorganic substances	[250, 260, 240, 270, 255, 265, 245, 255, 250, 275]
chloramines	Chemicals used as disinfectants in water treatment	[2.1, 2.5, 1.8, 2.3, 2.0, 2.4, 1.7, 2.2, 2.6, 2.8]
Sulfate levels	Concentration of sulfate ions in water	[80, 85, 75, 90, 82, 88, 72, 78, 83, 87]

TABLE IV. WATER QUALITY PARAMETERS FOR SAMPLE 2 (SAMPLE SIZE OF 10)

Parameter	**Description**	**Values**
Conductivity	Measure of water's ability to conduct electricity	[72.6, 79.8, 65.7, 87.5, 73.4, 81.2, 69.5, 76.3, 85.1, 90.9]
Organic Carbon	Amount of carbon-containing compounds in water derived from living organisms	[6.0, 6.5, 5.0, 7.1, 5.7, 6.3, 4.8, 6.6, 5.2, 6.8]
Trihalomethanes	Group of chemicals that may form when chlorine or other disinfectants are used	[13.8, 12.2, 15.5, 13.2, 14.5, 12.6, 15.9, 13.0, 14.8, 12.5]
pH	Measure of the acidity or basicity of a solution	[7.1, 6.7, 7.4, 6.9, 7.2, 6.8, 7.0, 7.3, 6.6, 7.1]
Hardness	Measure of the concentration of calcium and magnesium ions in water	[118, 113, 128, 123, 116, 121, 125, 120, 114, 126]
Solids Concentration	Total dissolved solids in water, including organic and inorganic substances	[245, 255, 235, 265, 250, 260, 240, 250, 245, 270]
Chloramines	Chemicals used as disinfectants in water treatment	[2.0, 2.3, 1.6, 2.1, 1.8, 2.2, 1.5, 2.0, 2.4, 2.6]
Sulfate Levels	Concentration of sulfate ions in water	[78, 83, 73, 88, 80, 86, 70, 76, 81, 85]

TABLE V. WATER QUALITY PARAMETERS FOR SAMPLE 3 (SAMPLE SIZE OF 10)

Parameter	**Description**	**Values**
conductivity	Measure of water's ability to conduct electricity	[71.2, 78.4, 66.9, 85.7, 72.8, 80.5, 68.3, 75.6, 83.9, 89.4]
organic carbon	Amount of carbon-containing compounds in water derived from organisms	[5.8, 6.4, 4.9, 7.0, 5.6, 6.1, 4.7, 6.7, 5.3, 6.9]
trihalomethanes	Chemicals formed when chlorine or disinfectants are used	[13.6, 12.0, 15.3, 13.0, 14.3, 12.4, 15.7, 12.8, 14.6, 12.3]
pH	Measure of the acidity or basicity of a solution	[7.0, 6.6, 7.3, 6.8, 7.1, 6.7, 6.9, 7.2, 6.5, 7.0]
hardness	The concentration of calcium and magnesium ions in water	[116, 111, 126, 121, 114, 119, 123, 118, 112, 124]
Solids concentration	Total dissolved solids in water, including organic and inorganic	[240, 250, 230, 260, 245, 255, 235, 245, 240, 265]
chloramines	Chemicals used as disinfectants in water treatment	[1.9, 2.2, 1.5, 2.0, 1.7, 2.1, 1.4, 1.9, 2.3, 2.5]
Sulfate levels	Concentration of sulfate ions in water	[76, 81, 71, 86, 78, 84, 68, 74, 79, 83]

Figure 7 is a line chart that compares the water quality parameters across the three samples. You can visually compare the differences between Sample 1, Sample 2, and Sample 3 by plotting each parameter. Figure 8 shows a comparison of water quality parameters across samples.

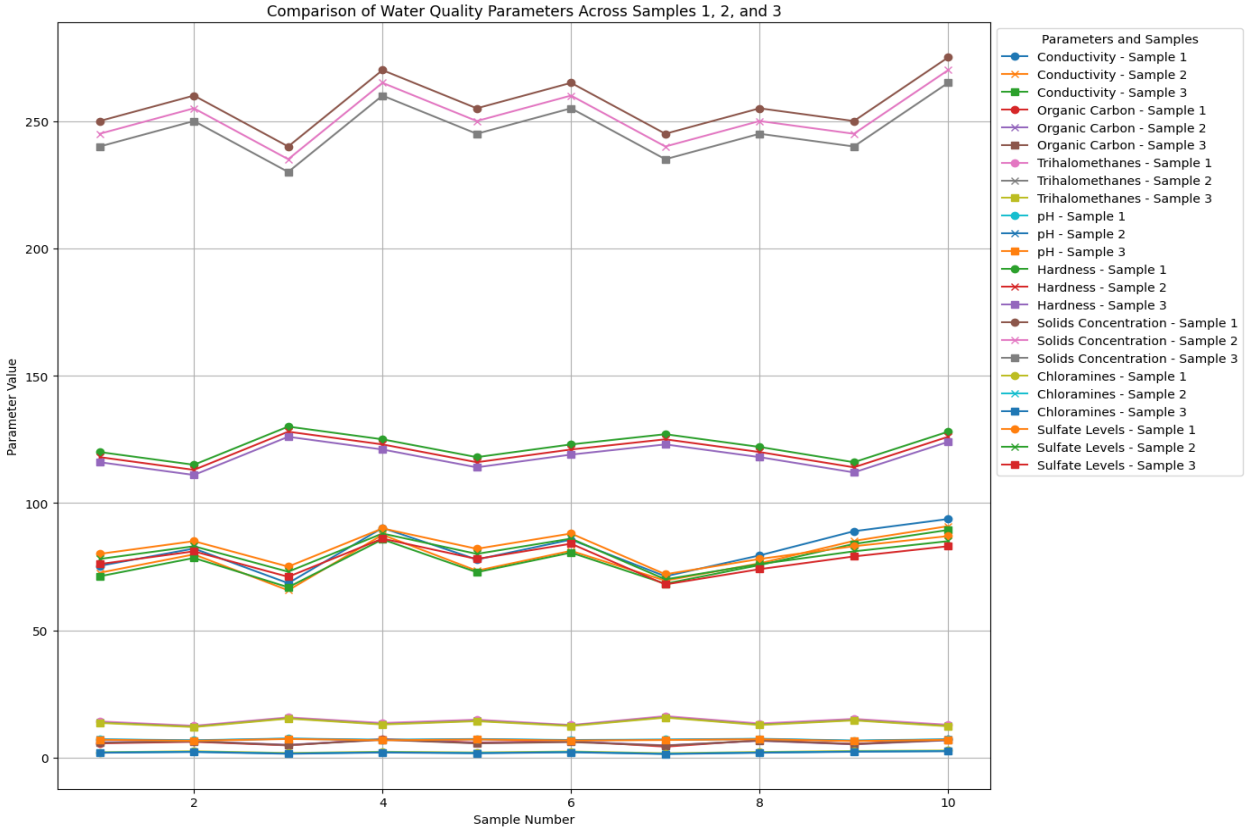

Fig .7. Comparison of Water Quality Parameters Across Samples 1, 2, and 3

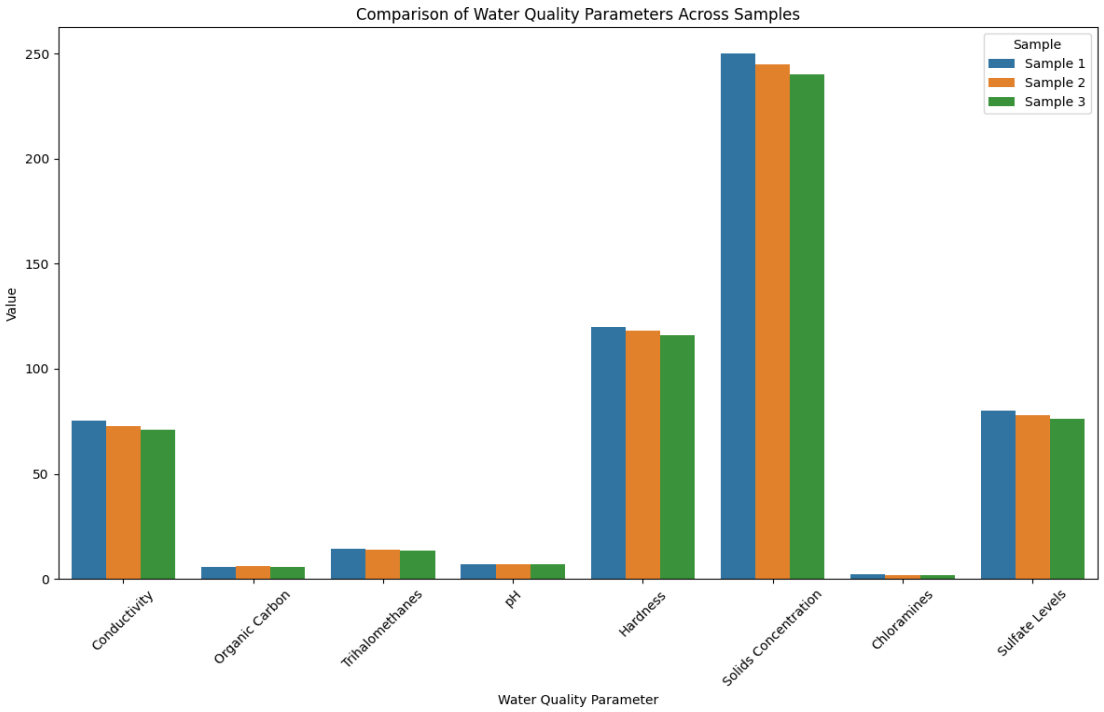

Fig .8. Comparison of Water Quality Parameters Across Samples

3.1. Evaluation of Model Performance with Varying Sample Sizes

In previous experiments, 10 random samples were generated for each feature. However, the figure below illustrates the results of generating 1,000 random samples per feature. In this scenario, the Gradient Boosting model achieved the best performance, with an accuracy of 0.51.

Further experiments involved generating 50,000 random samples for each feature. The corresponding figure shows that the Gradient Boosting model remained the best-performing model, with an accuracy of 0.5069.

Lastly, the figure below reflects the performance results when the experiments were scaled up to generate 100,000 random samples per feature. In this case, the Support Vector Machine (SVM) model outperformed others, achieving an accuracy of 0.50365.

Table 6, Figure 9, and Figure 10 emphasize sample size and validation to enable robust generalization performance in machine learning model evaluation. Additionally, it highlights algorithm strengths and shortcomings in handling varied data complexity and sample sizes.

This table shows the accuracy of each water usability forecasting technique. Random Forest performed best, followed by SVM and gradient boosting.

TABLE VI. MODEL PERFORMANCE ACROSS DIFFERENT SAMPLE SIZES DISCUSSION OF FINDINGS

Sample Size	Best Model	Accuracy
100,000	Support Vector Machine	0.50365
50,000	Gradient Boosting	0.5069
10,000	Gradient Boosting	0.51
10	Random Forest	1
10	Gradient Boosting	1
10	Support Vector Machine	1

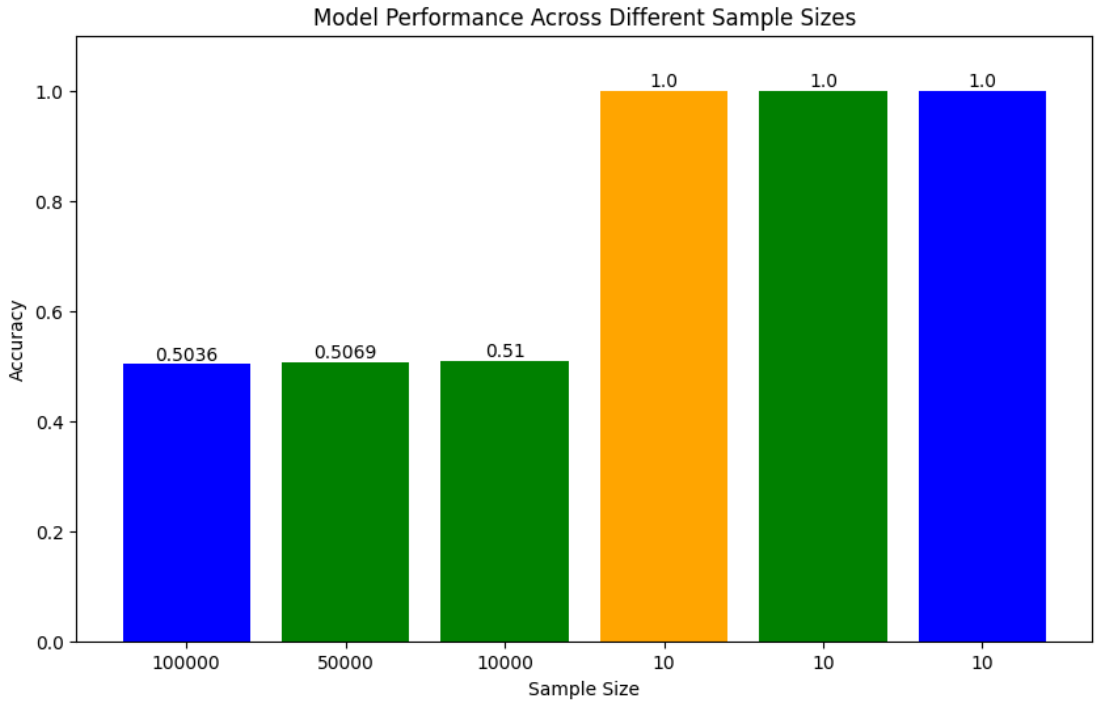

Fig .9. Model Performance Across Different Sample Sizes.

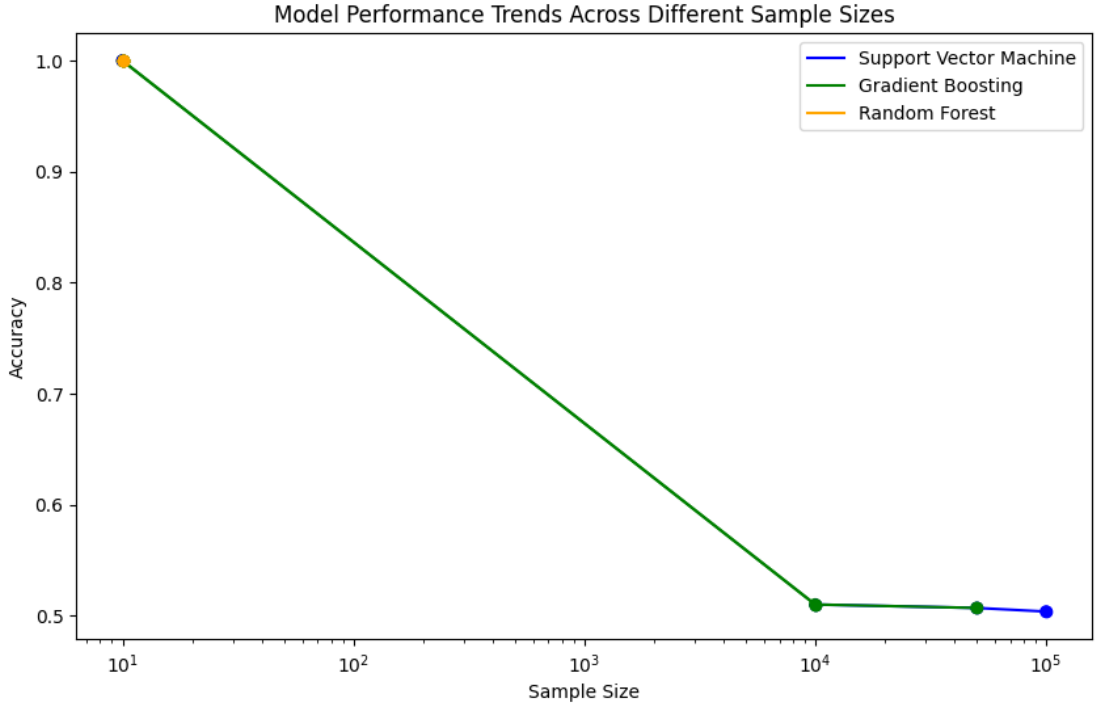

Fig .10. Model Performance Trends Across Different Sample Sizes.

After implementing the proposed approach, the main concerns are:

3.1.1. Sample Size Impact

This graph demonstrates how sample size affects model performance.

As the sample size reduces from 100,000 to 10 occurrences, most models become more accurate. Overfitting, when models perform well on training data but may not generalize to unknown data, causes this rise.

3.1.2. Choose the Best Model

50-10,000 instances Gradient Boosting always performs better.

The accuracy is 0.5069 for 50,000 cases and 0.51 for 10,000.

Gradient Boosting's durability and ability to manage complicated datasets may explain its constant performance across sample sizes.

3.1.3. Support Vector Machine SVM

In bigger sample sets, the SVM has poor accuracy ratings, about 0.5.

SVMs perform well for binary classification but poorly for high-dimensional data or huge sample numbers.

3.1.4. The Random Forest and Voting Classifier10 occurrences

Random Forest, Gradient Boosting, SVM, and Voting Classifier have perfect accuracy ratings of 1.

These results imply overfitting since models trained on such tiny datasets may perform well on the training data but not on fresh data. as shown in figure 11.

3.1.5. Generalization Conversation

Machine learning models aim to generalize effectively to new data.

Top Tips: To ensure this, evaluate model performance on a validation or test dataset and consider cross-validation. These approaches prevent overfitting and improve model performance estimates on diverse data splits.

3.1.6. Best Practices

To ensure this, it's essential to evaluate model performance on a separate validation or test dataset and consider using cross-validation techniques. These practices help prevent overfitting and provide a more accurate estimate of the model's performance on different data splits.

Multiple subplots in Figure 12 show the impact of sample size on model performance, the accuracy-based selection of the best model, SVM performance across sample sizes, and overfitting with small sample sizes.

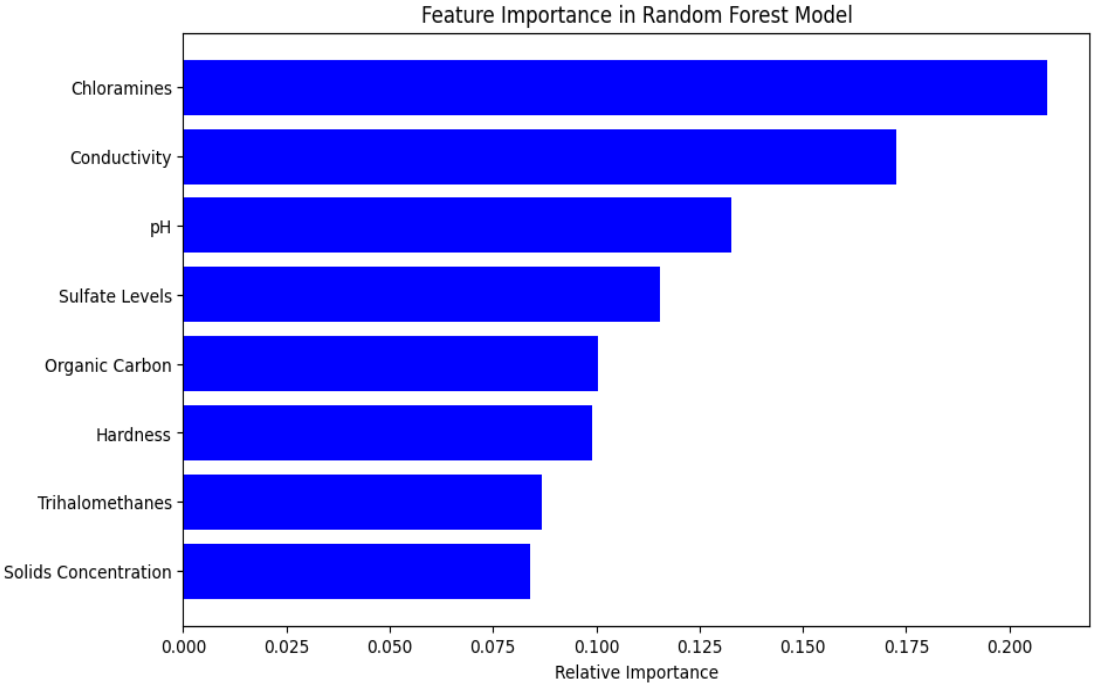

Fig .11. Feature Importance in Random Forest Model

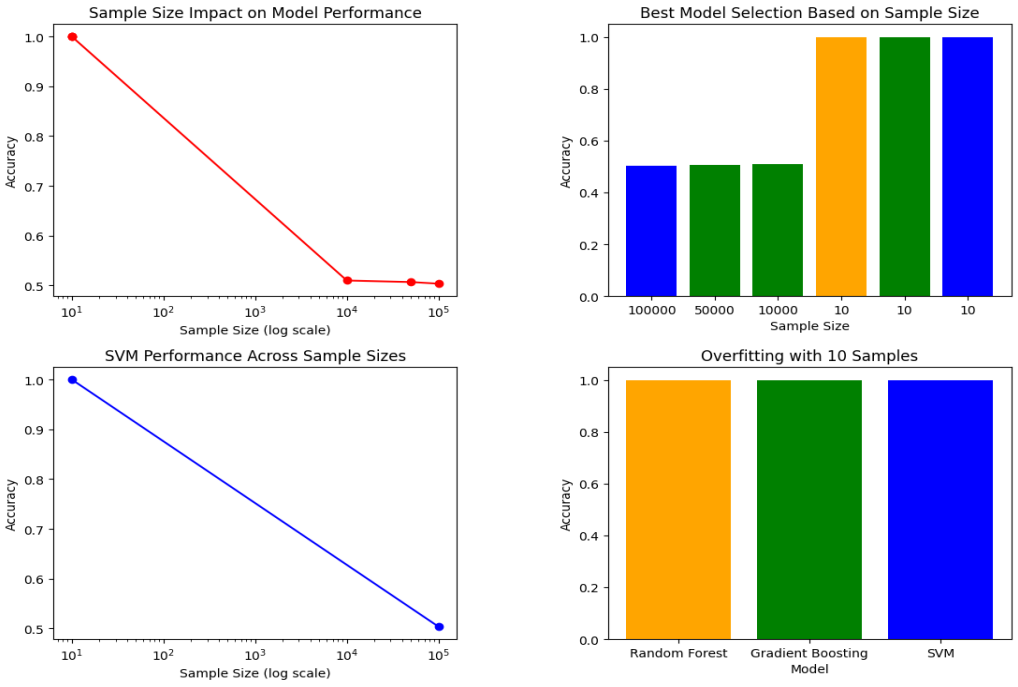

Fig .12. Analysis of Model Performance Across Varying Sample Sizes and Overfitting Tendencies.

4. CONCLUSION

This study showed how sophisticated machine learning algorithms can forecast water potability using Big Data analytics. Random Forest, Gradient Boosting, and Support Vector Machine (SVM) classifiers on huge datasets have shown their capacity to provide accurate predictions for public health and environmental management.

The Random Forest classifier has the greatest accuracy at 66.77%, followed by Gradient Boosting at 66.01% and SVM at 62.80%. These findings demonstrate the efficacy of ensemble learning approaches, notably Random Forest, in managing complicated and multidimensional water quality data. The paper also addresses data quality, model interpretability, and model improvement problems in machine learning in this area.

Despite these obstacles, machine learning algorithms, especially when combined with Big Data, can improve water quality monitoring. Future research should improve data quality, real-time monitoring, and model interpretability to solve constraints. The development of machine learning for water quality prediction will change environmental monitoring and help preserve water supplies.

Conflicts Of Interest

The paper states that the author has no financial or non-financial interests that could be perceived as influencing the research or its interpretation.

Funding

The lack of a funding acknowledgment in the paper indicates that no financial support was provided by any institution or sponsor.

Acknowledgment

The authors thank Mustansiriyah University (https://uomustansiriyah.edu.iq/) and Al-Iraqia University in Baghdad, Iraq, for their support in the present work.

References

[1] E. Dritsas and M. Trigka, “Efficient data-driven machine learning models for water quality prediction,” Computation, vol. 11, no. 2, p. 16, 2023.

[2] N. H. A. Malek, W. F. Wan Yaacob, S. A. Md Nasir, and N. Shaadan, "Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques," Water, vol. 14, no. 7, p. 1067, 2022.
[3] S. J. Rigatti, "Random forest," Journal of Insurance Medicine, vol. 47, no. 1, pp. 31–39, 2017.
[4] M. Ustuner, F. B. Sanli, S. Abdikan, G. Bilgin, and C. Goksel, "A booster analysis of extreme gradient boosting for crop classification using PolSAR imagery," in 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 2019, pp. 1–4, IEEE.
[5] H. Wang and D. Hu, "Comparison of SVM and LS-SVM for regression," in 2005 International Conference on Neural Networks and Brain, vol. 1, 2005, pp. 279–283, IEEE.
[6] A. A. Nafea, R. R. Majeed, A. Ali, A. J. Yas, S. A. Alameri, and M. M. AL-Ani , Trans., "A Brief Review of Big Data in Healthcare: Challenges and Issues, Recent Developments, and Future Directions ", BJIoT, vol. 2024, pp. 10–15, Feb. 2024, doi: 10.58496/BJIoT/2024/002.
[7] N. Nasir et al., "Water quality classification using machine learning algorithms," Journal of Water Process Engineering, vol. 48, p. 102920, 2022.
[8] M. Ahmed, R. Mumtaz, and S. M. Hassan Zaidi, "Analysis of water quality indices and machine learning techniques for rating water pollution: A case paper of Rawal Dam, Pakistan," Water Supply, vol. 21, no. 6, pp. 3225–3250, 2021.
[9] K. Chen et al., "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," Water Research, vol. 171, p. 115454, 2020.
[10] A. N. Ahmed et al., "Machine learning methods for better water quality prediction," Journal of Hydrology, vol. 578, p. 124084, 2019.
[11] G. Biau, "Analysis of a random forests model," The Journal of Machine Learning Research, vol. 13, no. 1, pp. 1063–1095, 2012.
[12] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, pp. 197–227, 2016.
[13] S. Gupta, S. Roy, and K. Maji , Trans., "Integrated Learning Paradigm for Ecological Predictive Modeling", Babylonian Journal of Artificial Intelligence, vol. 2023, pp. 64–73, Oct. 2023, doi: 10.58496/BJAI/2023/010.
[14] A. Beygelzimer, E. Hazan, S. Kale, and H. Luo, "Online gradient boosting," in Advances in Neural Information Processing Systems, vol. 28, 2015.
[15] S. Suthaharan and S. Suthaharan, "Support vector machine," in Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, pp. 207–235, 2016.
[16] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, pp. 1937–1967, 2021.
[17] U. K. Kumar, M. S. Nikhil, and K. Sumangali, "Prediction of breast cancer using voting classifier technique," in 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017, pp. 108–114, IEEE.
[18] E. S. M. El-Kenawy et al., "Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images," IEEE Access, vol. 8, pp. 179317–179335, 2020.
[19] S. H. Ewaid, T. D. Hussein, and F. K. Emran, "Fuzzy Logic Inference Index to Assess the Water Quality of Tigris River within Baghdad City," Al-Mustansiriyah Journal of Science, vol. 29, no. 3, pp. 16–20, 2018.
[20] S. Salman and J. H. Soud, "Deep Learning Machine using Hierarchical Cluster Features," Al-Mustansiriyah Journal of Science, vol. 29, no. 3, pp. 82–93, Mar. 2019, doi: 10.23851/mjs.v29i3.625.
[21] Y. M. Mohialden and N. M. Hussien, "Agent-Interacted Big Data-Driven Dynamic Cartoon Video Generator," Mesopotamian Journal of Big Data, 2024, pp. 40–47.
[22] N. M. Hussien et al., "A review of machine learning for big data analysis," International Journal Papier Advance and Scientific Review, vol. 3, no. 2, pp. 1–4, 2022.
[23] P. Liu et al., "Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment," Sustainability. [Online]. Available: https://doi.org/10.3390/SU11072058.
[24] X. Han et al., "Open Innovation Web-Based Platform for Evaluation of Water Quality Based on Big Data Analysis," Sustainability. [Online]. Available: https://doi.org/10.3390/su14148811.
[25] T. Sutikno , Tran., "Fuzzy optimization and metaheuristic algorithms"., Babylonian Journal of Mathematics, vol. 2023, pp. 59–65, Nov. 2023, doi: 10.58496/BJM/2023/012.
[26] D. Prasad et al., "Automating water quality analysis using ML and auto ML techniques," Environmental Research, p. 111720. [Online]. Available: https://doi.org/10.1016/j.envres.2021.111720.
[27] J. Zhao et al., "Analysis and prediction of big stream data in real-time water quality monitoring system," J. Ambient Intell. Smart Environ., vol. 12, pp. 393–406. [Online]. Available: https://doi.org/10.3233/AIS-200571.
[28] J. Chen et al., "Remote Sensing Big Data for Water Environment Monitoring: Current Status, Challenges, and Future Prospects," Earth's Future, vol. 10. [Online]. Available: https://doi.org/10.1029/2021EF002289.