Hydrology Research

Hydrology Research Vol 55 No 7, 775 doi: 10.2166/nh.2024.048

Water quality level estimation using IoT sensors and probabilistic machine learning model

Mahesh T.R.[a], Surbhi Bhatia Khan[b,c,*], A. Balajee[a], Ahlam Almusharraf[d], Thippa Reddy Gadekallu[e,f], Eid Albalawi[g] and Vinoth Kumar V.[h]

[a] Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru 562112, India
[b] School of Science, Engineering and Environment, University of Salford, M5 4WT Manchester, UK
[c] Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon, Lebanon
[d] Department of Business Administration, College of Business and Administration, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671 Riyadh, Saudi Arabia
[e] Division of Research and Development, Lovely Professional University, Phagwara, India
[f] Center of Research Impact and Outcome, Chitkara University, Punjab, India
[g] Department of Computer science, College of Computer Science and Information Technology, King faisal University, 31982 Hofuf, Saudi Arabia
[h] School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore 632001, India
*Corresponding author. E-mail: s.khan138@salford.ac.uk

ABSTRACT

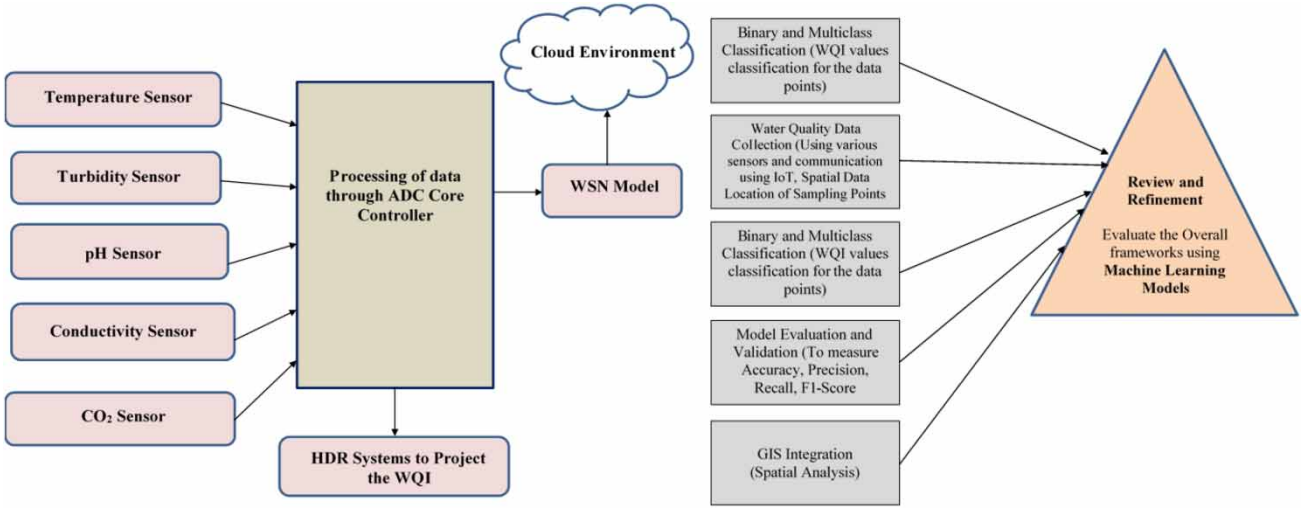
Drinking water purity analysis is an essential framework that demands several real-world parameters to ensure the quality of water. So far, sensor-based analysis of water quality in specific environments is done concerning certain parameters including the PH level, hardness, TDS, etc. The outcome of such methods analyzes whether the environment provides potable water or not. Potable denotes the purified water that is free from all contaminations. This analysis gives an absolute solution whereas the demand for drinking water is a growing problem where the multiple-level estimations are essential to use the available water resources efficiently. In this article, we used a benchmark water quality assessment dataset for analysis. To perform a level assessment, we computed three major features namely correlation-entropy, dynamic scaling, and estimation levels, and annexed with the earlier feature vector. The assessment of the available data was performed using the statistical machine learning model that ensembles the random forest model and light gradient boost model (GBM). The probability of the ensemble model was done by the Kullback Libeler Divergence model. The proposed probabilistic model has achieved an accuracy of 96.8%, a sensitivity of 94.55%, and a specificity of 98.29%.

Key words: correlation-entropy, dynamic scaling, estimation, gradient boost model, linear discriminant analysis (LDA)

HIGHLIGHT

- Developing an IoT framework for water quality management data extraction involves deploying a network of sensors capable of measuring key parameters such as pH, dissolved oxygen, turbidity, temperature, conductivity, and contaminants across water bodies. The proposed work uses a probabilistic machine learning model to estimate the multiple levels of water quality assessments.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Water as an essential need for every human being around the globe demands a systematic framework for analyzing its potable nature and also the parameters to ensure its integrity. Numerous parameters are used to represent the integrity of water including pH levels, saltation levels, mineral levels, TDS, etc. (Qi *et al.* 2022). These are biological parameters that represent the quality of water management. The water quality index (WQI) is constructed by integrating various water quality parameters into a single numerical value, providing a comprehensive assessment of overall water quality. Initially, a selection of relevant parameters such as pH, dissolved oxygen, turbidity, temperature, conductivity, and levels of contaminants are identified based on their significance in determining water quality (Gaur *et al.* 2022). These parameters are assigned weights or importance factors reflecting their relative significance in influencing water quality. Subsequently, individual parameter values are normalized and transformed into dimensionless scores using appropriate mathematical functions or scales to ensure comparability across different parameters (Uddin *et al.* 2023). These normalized scores are then aggregated using weighted averaging or another aggregation method to compute a single composite score, representing the overall water quality status. The final step involves categorizing the composite score into predefined quality classes or categories (e.g., excellent, good, fair, poor) to facilitate interpretation and decision-making. The WQI framework provides a simplified yet informative way to assess and communicate water quality status to stakeholders, enabling informed management decisions and interventions aimed at preserving and improving water resources (Uddin *et al.* 2021). With the biological parameters, only the assessment regarding a specific environment can be performed whereas the computational exploration of those parameters could provide a global framework for analyzing the water quality (Liu *et al.* 2022).

The computational exploration can be done using different strategies and the data extraction can be done using the IoT sensors that transfer the obtained data and the parameter analysis can be performed based on the nature of the parameters (Cao *et al.* 2022). Water quality assessment through IoT devices involves the strategic deployment of sensors capable of measuring various parameters such as pH, dissolved oxygen, turbidity, temperature, conductivity, and contaminant levels within water bodies. Initially, appropriate sensors are selected based on the specific parameters requiring monitoring. These sensors are then integrated into IoT devices, typically microcontrollers or single-board computers like Raspberry Pi, equipped with communication modules such as Wi-Fi, Bluetooth, or cellular connectivity (Kruse 2018). The sensors are strategically deployed at desired locations within the water bodies, ensuring proper calibration and positioning for accurate measurements. The IoT devices collect data from the sensors at regular intervals, with the frequency of data collection tailored to the monitoring requirements. Subsequently, the collected data are transmitted wirelessly to a central server or cloud platform using the communication capabilities of the IoT devices (Jayaraman *et al.* 2024). Upon transmission, the data are securely stored in a database or cloud storage, where it undergoes analysis to identify patterns, trends, and anomalies. Visualization tools and dashboards are developed to provide real-time insights into water quality conditions, enabling

stakeholders to make informed decisions. Regular maintenance and calibration of sensors are conducted to ensure measurement accuracy, with malfunctioning components promptly replaced or repaired. Integration with decision support systems further enhances the utility of the data by linking it with predictive models or resource management systems, facilitating timely interventions to address any detected issues. Overall, water quality assessment via IoT devices enables continuous monitoring, proactive management, and effective preservation of water resources. The iterative modeling of water quality assessment can be carried out by inducing machine learning models to analyze the water quality-related data (Bedi *et al.* 2020). Water quality assessment employing machine learning models entails the utilization of historical water quality data to develop predictive algorithms capable of analyzing and forecasting water quality parameters. Initially, a comprehensive dataset containing information on various water quality parameters, alongside relevant environmental factors such as weather patterns, land use, and hydrological characteristics, is compiled (Thorslund & van Vliet 2020). This dataset serves as the foundation for training machine learning models, including regression, classification, and clustering algorithms, among others. The models are trained to recognize patterns, correlations, and anomalies within the data, enabling them to predict water quality parameters based on input variables. Once trained, these models can analyze real-time data collected from sensors deployed in water bodies, providing continuous assessment and early detection of water quality fluctuations. Additionally, machine learning models can be utilized to optimize sampling strategies, identify sources of contamination, and prioritize management interventions, thereby enhancing the efficiency and efficacy of water quality monitoring and management efforts. Regular updates and refinements to the models ensure their adaptability to changing environmental conditions and evolving water quality challenges, ultimately contributing to the sustainable management and preservation of water resources. Utilizing machine learning models for water quality assessment involves extracting pertinent features from sensor-gathered raw data in water bodies. These features are then utilized to train predictive algorithms. Initially, a comprehensive dataset comprising raw measurements of water quality parameters (e.g., pH, dissolved oxygen, turbidity, temperature, and conductivity) is compiled. Domain knowledge and statistical techniques are subsequently employed to compute additional features, including statistical moments, frequency-domain attributes, time series characteristics, and spatial patterns from the raw data. These computed features act as inputs for machine learning algorithms, facilitating the learning of intricate relationships between input features and target water quality parameters. Techniques for feature selection may also be applied to pinpoint the most informative features, thus enhancing the predictive performance of the models while reducing dimensionality and computational complexity. Post-training, these machine learning models can analyze real-time sensor data to predict water quality parameters, furnishing valuable insights into the current state of water quality and enabling timely management interventions. Continuous refinement and adaptation of feature computation techniques and machine learning models are crucial for effectively addressing evolving water quality challenges and enhancing water quality assessment systems' overall accuracy and reliability. There are three major objectives proposed in this article for estimating the water quality using IoT, and a machine learning model as follows:

- IoT sensor-based data collection model from the real-world environment
- A triple-stage feature computation model to increase the size of the feature vector
- Probabilistic machine learning model to estimate the multiple levels of water quality assessments.

Section 2 denotes the literature review followed by Section 3 which denotes the materials and methods that are adopted for the implementation Section 4 which illustrates the experimental results that are obtained through the proposed model and discusses the obtained results followed by the conclusion and references.

2. LITERATURE SURVEY

The literature review on water quality assessment begins with analyzing the physical changes that are made in water quality data extraction where the parameters are extracted using the traditional methods where the purity analysis using pH calculation is performed initially (Wu *et al.* 2021). Because of its broad structure, the assessment highlights the consolidation of extensive water quality data into a singular value or index as a valuable method. This process involves four sequential stages within WQI models: initial selection of water parameters for sub-index generation, assigning weights to these parameters, and calculating the overall WQI. Consequently, a vast amount of water quality data is condensed into a solitary index, allowing for comparison among different traditional WQI indices concerning parameter selection, sub-index creation, weight assignment, aggregation methods, and rating scales (Nayak *et al.* 2020). After evaluating the WQI, it was determined that traditional calculations consumed excessive time and revealed a limited number of errors during the sub-index

calculations. Various statistical and visual evaluation indicators were employed to assess the models. The data underwent division into training and testing sets by the machine learning algorithm, which utilized hybrid algorithms and provided estimations for WQI values (Bui *et al.* 2020). The research conducted for WQI analyses took place in Lake Poyang, China, involving the classification of 24 water quality samples into three distinct groups. The analysis focused on 20 diverse water quality parameters, particularly emphasizing Total Nitrogen (TN) and Total Phosphorus (TP), while assigning lower ratings to hazardous metals and other criteria in the WQI Analysis (Wu *et al.* 2017). It engaged in the utilization of a comprehensive range of fractional deviation techniques encompassing difference, ratio, and normalized difference indices. The resultant WQI values exhibit a range from 56.61 to 2,886.51. The index above was derived through the assessment of curve slope and root mean square error values (Wang *et al.* 2017). A variety of feature extraction methods are employed in water quality assessment to distill relevant information from raw data collected by sensors and other monitoring devices (Sagan *et al.* 2020). One prevalent approach involves statistical methods such as mean, median, standard deviation, and skewness, which provide insights into the central tendency, variability, and distribution of water quality parameters (Kumar & Padhy 2014). Time-domain analysis techniques, including autocorrelation and spectral analysis, capture temporal patterns, and periodicities in water quality data, aiding in the identification of seasonal variations and long-term trends (Akbarighatar *et al.* 2023). Frequency-domain methods, such as Fourier analysis and wavelet transforms, enable the decomposition of signals into frequency components, facilitating the detection of cyclic patterns and oscillations (Condon *et al.* 2021). Spatial analysis techniques, such as kriging and spatial interpolation, are utilized to analyze spatial variability and spatial autocorrelation in water quality parameters across different locations within a water body (Monica & Choi 2016). Additionally, machine learning algorithms such as principal component analysis (PCA), independent component analysis (ICA), and feature selection algorithms like genetic algorithms and recursive feature elimination are employed to identify and prioritize the most informative features for water quality assessment (Hosseini Baghanam *et al.* 2022). These feature extraction methods collectively contribute to a comprehensive understanding of water quality dynamics, aiding in the detection of anomalies, prediction of future trends, and formulation of effective management strategies for preserving and improving water resources (Dilmi & Ladjal 2021). A diverse array of machine learning classification methods is employed in water quality assessment to effectively analyze and categorize water quality data. Supervised learning algorithms such as decision trees, random forests, and support vector machines (SVMs) are widely utilized to classify water samples into different quality categories based on their feature vectors (Najwa Mohd Rizal *et al.* 2022). These algorithms leverage labeled training data to learn decision boundaries and classify unseen samples with high accuracy. Additionally, ensemble methods like AdaBoost and gradient boosting combine multiple weak classifiers to enhance classification performance (Khan *et al.* 2022). Deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are increasingly employed to extract intricate patterns and temporal dependencies from water quality data, particularly in spatial and temporal modeling tasks (Baek *et al.* 2020). Unsupervised learning methods such as clustering algorithms, including k-means and hierarchical clustering, are utilized for exploratory analysis and pattern recognition in unlabeled data, enabling the identification of natural groupings and anomalies in water quality datasets (Marin Celestino *et al.* 2018). Furthermore, hybrid approaches that integrate multiple machine learning techniques, such as feature selection, dimensionality reduction, and ensemble learning, are employed to improve classification accuracy and robustness in complex water quality assessment tasks (Aslam *et al.* 2022). Overall, machine learning classification methods play a pivotal role in effectively analyzing and interpreting water quality data, facilitating informed decision-making and management of water resources.

3. MATERIALS AND METHODS

We proposed three different strategies for water quality assessment and created three major estimation levels from the global dataset. Initially, we proposed an IoT framework to extract real-world data for water quality assessment. Once the data are retrieved, we computed three major features that are essential to discriminate the different levels of estimation for the water quality assessment. Once the feature vector for water quality assessment was created, we proposed a probabilistic machine learning model to perform multiclass classification to match the different estimation levels of water quality management. The overall workflow of the proposed method has been given in Figure 1.

3.1. Dataset description

We used the global water quality dataset retrieved from the benchmark Kaggle repository for implementing the proposed methods. The dataset consists of 3,277 instances with 10 major features, namely pH level of the water, hardness, solids,

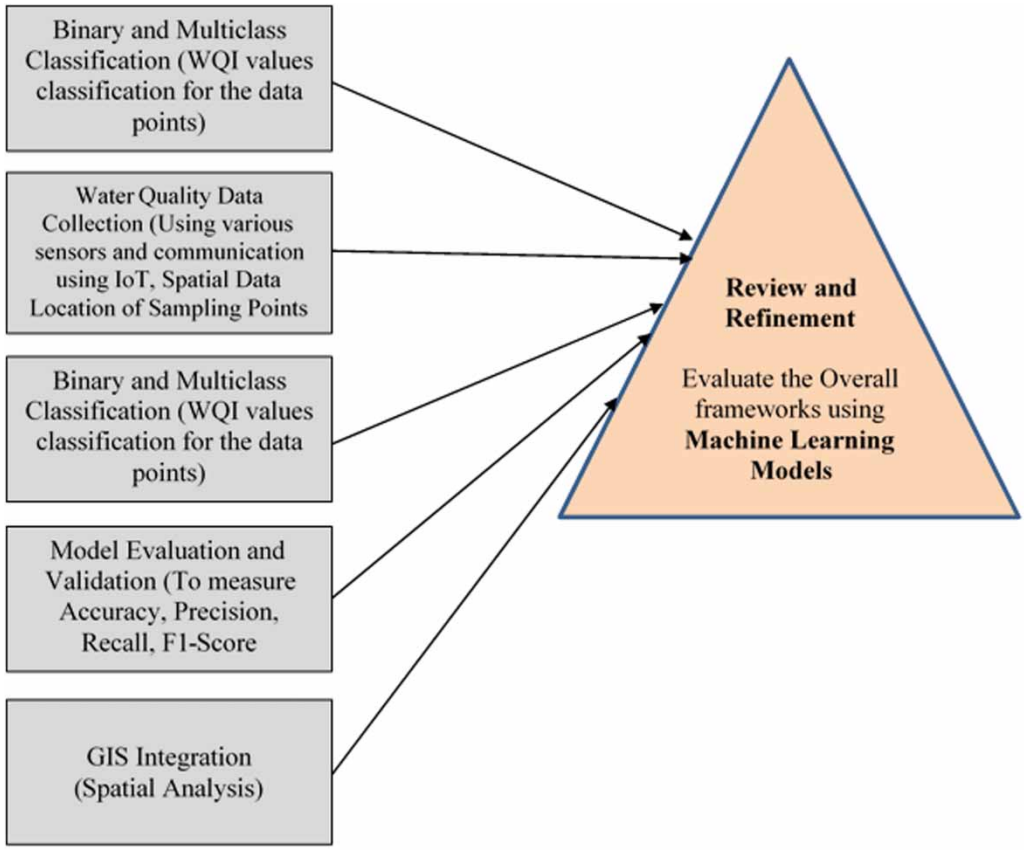


Figure 1 | Workflow of the proposed model.

amount of chloramines, amount of sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potable nature of the water. The only class description in the obtained dataset is the potable nature whereas all the other features remain as problem-related features.

3.2. IoT framework to extract the water quality management data

Developing an IoT framework for water quality management data extraction involves deploying a network of sensors capable of measuring key parameters such as pH, dissolved oxygen, turbidity, temperature, conductivity, and contaminants across water bodies. These sensors are connected to IoT devices equipped with communication modules like Wi-Fi, cellular, or LoRa, facilitating data collection and transmission to a central server or cloud platform (Liu *et al.* 2023; Zhou *et al.* 2024). Firmware/software in IoT devices collects data at regular intervals, ensuring real-time monitoring. Data are stored securely in databases or data lakes, enabling efficient management and analysis.

Interactive dashboards and mobile apps provide stakeholders with real-time visualization of water quality parameters, facilitating informed decision-making (Chaudhari 2019). Machine learning algorithms analyze data to detect patterns, anomalies, and trends, supporting predictive modeling and decision support (Dai *et al.* 2024a, 2024b). Integration with decision support systems and regulatory compliance tools ensures seamless data exchange and regulatory adherence (Ahmed *et al.* 2019). The framework prioritizes scalability, flexibility, and security, allowing for easy expansion, interoperability with different sensors, and robust protection against cyber threats. Regular maintenance and updates ensure accurate and reliable data collection, contributing to proactive water quality management and resource preservation (Zhang *et al.* 2021). Figure 2 denotes the IoT framework to extract the data from the real-world environment. Here CO_2 sensors can play a crucial role in water quality assessment by detecting carbon dioxide levels dissolved in water, which can indicate the presence of organic matter decomposition, industrial pollution, or

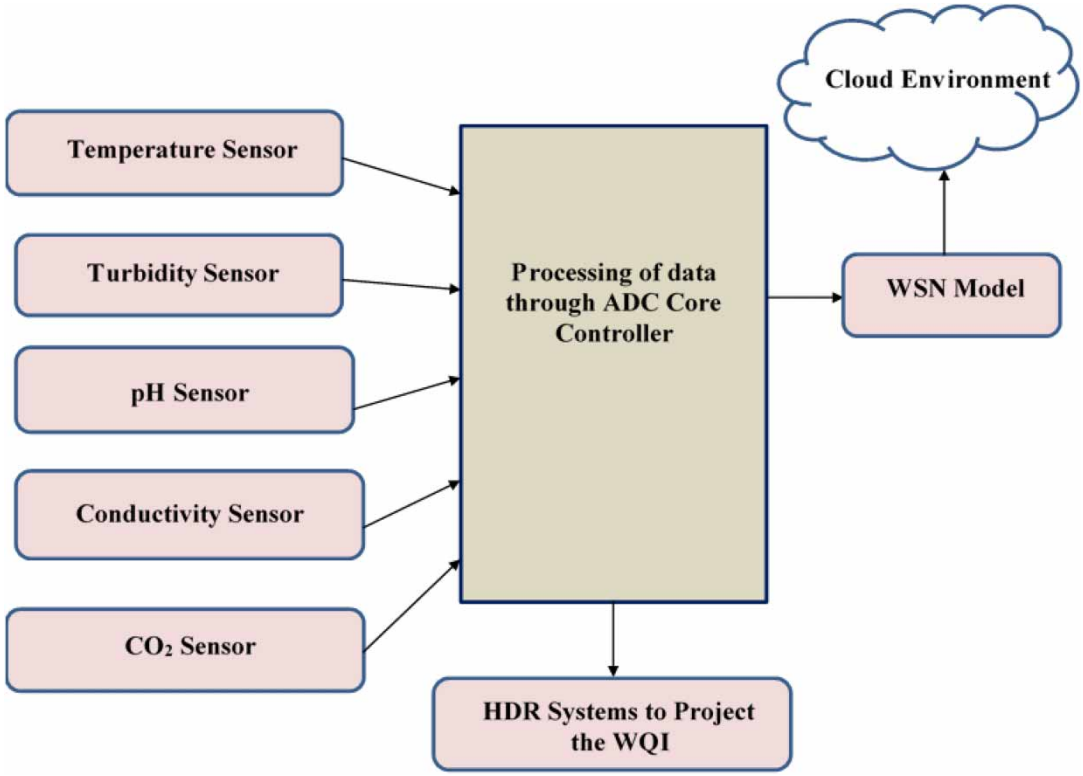


Figure 2 | IoT framework proposed for water quality data collection.

other environmental factors affecting water quality (Luo *et al.* 2022). These sensors typically employ a membrane-based or solid-state electrochemical sensing mechanism. In membrane-based sensors, a semi-permeable membrane allows only CO_2 molecules to diffuse through, where they interact with a pH-sensitive indicator solution, leading to a change in pH that is measured by a pH electrode (Konde & Deosarkar 2020). Solid-state electrochemical CO_2 sensors utilize electrodes and an electrolyte to produce a voltage proportional to the concentration of CO_2 in water. The measured CO_2 concentration can provide insights into the biological activity, acidity, and overall health of aquatic ecosystems, aiding in the early detection of pollution events and informing management decisions for water resource conservation and protection.

3.3. Triple-stage feature extraction model

As mentioned in the dataset description, the available dataset has only a single estimation feature, i.e. the potable nature of the water to analyze whether it attained the quality or not. The major objective is to frame the estimation levels so that further contamination levels can be eliminated from the analysis and the remaining levels are tuned for the purification process. The triple-stage feature extraction begins with creating a multidimensional scaling as a feature that can be computed by representing the available features based on their similarity and dissimilarity matrix. Let D be the dissimilarity matrix computed from the available samples, where D_{ij} represents the dissimilarity or distance between samples i and j. MDS aims to find a low-dimensional representation of these samples while preserving their pairwise distances as much as possible. Let X be the matrix representing the low-dimensional embedding of the samples, where each row x_i corresponds to the coordinates of the sample i in the reduced space. We aim to minimize the stress function, which measures the discrepancy between the original dissimilarities and the distances in the embedded space:

$$\text{Stress} = \sum_{i<j} (d_{ij} - \|x_i - x_j\|)^2 \quad (1)$$

where d_{ij} is the dissimilarity between samples i and j and $\|x_i - x_j\|$ is the Euclidean distance between their coordinates in the reduced space. The coordinates x_i can be obtained by solving the optimization problem which denotes the MDS equation from the obtained matrix.

$$\text{MDS} = \min(X) \sum_{i<j} (D_{ij} - \|x_i - x_j\|)^2 \tag{2}$$

Equation (2) is subject to constraints or using iterative algorithms such as gradient descent, classical MDS, or Sammon mapping (Tang *et al.* 2020). Once the coordinates x_i are obtained, they can be used as features in the dataset for further analysis or modeling. Once the scaling is done, then the next stage is to identify the correlation within the available features and to form a normalized value as one of the feature sections. This was done by computing Kendall's rank correlation coefficient. Kendall's rank correlation coefficient, also known as Kendall's tau (τ), is a nonparametric measure of association between two variables based on the ranks of the data. Let X and Y be two variables for which we want to compute Kendall's tau. The formula for Kendall's tau (τ) can be represented mathematically as follows:

$$\tau = \frac{\sum_{i<j} \text{sign}((x_i - x_j).(y_i - y_j))}{\sqrt{\frac{(2.np)}{n(n-1)}}} \tag{3}$$

Here x_i and y_i denote the ranks of the ith observation in variables X and Y, respectively. n is the number of observations. A sign is the sign function, returning $+1$ if its argument is positive, -1 if negative, and 0 if zero. The summation is over all distinct pairs $i < j$ of observations. Kendall rank correlation offers several advantages as a feature extraction model. Firstly, it is a nonparametric measure, meaning it does not rely on assumptions about the underlying distribution of the data, making it robust to outliers and nonnormality (Chen *et al.* 2022). Secondly, Kendall's tau is invariant to monotonic transformations of the data, allowing it to capture nonlinear relationships between variables. Additionally, Kendall's tau is less affected by tied ranks compared to other correlation measures, such as Pearson correlation, making it suitable for datasets with tied observations. Furthermore, Kendall's tau is interpretable, ranging from -1 to $+1$, with higher absolute values indicating stronger associations between variables (Pandey *et al.* 2018). Overall, Kendall rank correlation provides a flexible and robust approach for extracting features from data, particularly in scenarios where nonlinearity, nonnormality, and tied ranks are prevalent. Since in water quality assessment, there exists a nonlinear nature where the data keep on fluctuating in terms of time constraint, the Kendall rank correlation suits as a feature for the dataset considered for analysis. The third stage in the proposed feature extraction process is to compute a categorical value as a feature. We adopted a categorical encoding method to compute the feature. The feature named 'Estimation Level' is computed through the categorical encoding method. Categorical encoding is the method of converting the categorical variables into a numerical representation for better analysis (Cerda *et al.* 2018). This process begins with the assumption C be a categorical variable with m unique categories. The encoded representation is a vector V_i which can be framed through Equation (2).

$$V_i[j] = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Using the categorical encoding, the estimation levels have been framed into three categories, namely Excellent, Good, and Poor. Categorical encoding offers several advantages as a feature extraction model. Firstly, it allows for the representation of categorical data in a format that can be understood and utilized by machine learning algorithms, enabling the inclusion of categorical variables in predictive models. Secondly, categorical encoding techniques such as one-hot encoding and ordinal encoding preserve the inherent information and relationships within categorical variables, ensuring that important distinctions between categories are retained during feature extraction. The proposed triple-stage feature extraction algorithm is given below.

Algorithm 1: Triple Stage Feature Extraction	
Input	Water quality samples $i, i_2, i_3 \ldots .. i_n$ feature collections $f_1, f_2, f_3 \ldots .. f_n$
Output	An extended feature set with problem-related features
Begin	
Step 1:	For every attribute f_i in feature matrix f
Step 2:	Perform combinational analysis f(x)
Step 3:	Calculate the dissimilarity using Equation (1)
Step 4:	For the individual parameters i and j.
Step 5:	Calculate the Euclidean distance $\|x_i - x_j\|$
Step 6:	Update the feature vector with $MDS = \min(X) \sum_{i<j} (D_{ij} - \|x_i - x_j\|)^2$
Step 7:	Calculate the correlation between all i and j
Step 8:	*if* i and j are non-parametric
Step 9:	Calculate $\tau = \frac{\sum_{i<j} sign((x_i - x_j).(y_i - y_j))}{\sqrt{\frac{(2.np)}{n(n-1)}}}$ and add to feature vector
Step 10:	Perform categorical encoding using V_i and V_j
Step 11:	End if
Step 12:	End for
Step 13:	Repeat steps 2 to 10 to increase the size of the feature vector
Step 14:	Return the feature vector
End	

Additionally, categorical encoding facilitates the incorporation of domain knowledge and prior information about the categorical variables into the feature space, enhancing the interpretability and explainability of the resulting models. Furthermore, categorical encoding can effectively handle nominal and ordinal categorical variables, accommodating a wide range of categorical data types commonly encountered in real-world datasets. The proposed feature handling mechanism not only initiates the computation of problem-related features but a filtering with respect to the missing values is also taken care of during the preprocessing stage. In the preprocessing stage, the proposed mechanism eradicates the missing values by having a threshold value of 0.5 which was 5% per variable in the set. Once the categorical features are computed the size of the feature vector has 13 features and 3,277 instances.

3.4. Probabilistic gradient boost model for multiclass classification

The feature vector obtained after triple-stage feature extraction was used for the analyzing the performance of the proposed classification model. The original dataset had the potable feature which had a balanced data distribution whereas the computed categorical feature had four different classes and there was an imbalance in the dataset. To balance the feature vector, the synthetic minority oversampling technique (SMOTE) was used. SMOTE stands as a prominent algorithm in imbalanced learning, tailored specifically to counter the challenge of class imbalance in datasets. Its core function lies in generating synthetic samples for the minority class, thus achieving a balanced class distribution. Initially, SMOTE identifies minority class instances that closely align in the feature space, creating synthetic instances along the line segments connecting these instances. This method entails randomly selecting a minority class instance and one of its nearest neighbors, thereafter generating a synthetic sample along the joining line segment. This approach effectively augments the minority class size, mitigating the class imbalance issue. Nonetheless, it's noteworthy that SMOTE may introduce noise into the dataset, particularly in cases where the minority class heavily overlaps with the majority class. In response, variants like Borderline-SMOTE and ADASYN have emerged, refining SMOTE's performance by targeting borderline instances or adapting the synthetic sample generation process based on local class distribution. Overall, SMOTE emerges as a potent tool for tackling class imbalance, elevating the efficacy of machine learning models in imbalanced datasets. The SMOTE analysis created a

balanced dataset with 655 normal samples, 655 samples belonging to classes 1 to 3, and 657 samples from class 4 where the earlier sample sets had only 167 samples belonging to the normal state. The construction of the proposed model begins with training the model using the gradient boost machines (GBM). Gradient boosting is a machine learning technique that builds an ensemble of weak learners, typically decision trees, sequentially, with each tree learning to correct the errors of the previous ones. Let $(x_i,\ y_i)$ represent the ith training example, where x_i is the feature vector and y_i is the corresponding target variable. The objective function assigned by the GBM model can be denoted using the following equation.

$$\text{Obj} = \sum_{i=1}^{N} L(y_i,\ F(x_i)) + \sum_{m=1}^{M} \Omega(f_m) \tag{5}$$

The above equation represents the objective function, which combines the loss function and a regularization term $\Omega(f_m)$ to prevent overfitting. GBMs have revolutionized machine learning by offering a potent ensemble learning technique ideal for diverse predictive modeling tasks. GBMs construct highly accurate predictive models by sequentially combining numerous weak learners, predominantly decision trees, to mitigate individual model limitations and enhance predictive performance. Their prowess extends to handling intricate, nonlinear data relationships, rendering them invaluable for regression, classification, and ranking tasks. GBMs demonstrate remarkable flexibility, adeptly adapting to diverse data types and accommodating various feature representations, thus finding extensive application across real-world scenarios. Additionally, GBMs ensure robustness against overfitting through the integration of regularization techniques and adaptive learning rates, facilitating effective generalization to unseen data. Notably, GBMs offer interpretability, furnishing insights into feature importance and model behavior, thereby aiding comprehension of underlying data patterns. With the evolution of implementation frameworks like XGBoost, LightGBM, and CatBoost, GBMs have become indispensable tools for data scientists, propelling innovations in predictive modeling, recommendation systems, anomaly detection, and beyond. Ultimately, GBMs' capacity to deliver superior predictive accuracy, flexibility, robustness, interpretability, and scalability solidifies their status as fundamental assets in the machine learning arsenal, proficiently addressing an extensive array of challenges. The trained GBM model was executed as a single-phase decision tree model and the initial validation is done. Once the results of single DT are analyzed an ensemble of DT, i.e. random forest model, is chosen for executing the GBM algorithm for all the sets available in the feature vector. Once the random outcomes are obtained, the majority voting is performed by the probabilistic Kullback–Libeler Divergence model. The significance of Kullback–Leibler (KL) divergence in multiclass classification lies in its ability to quantify the difference between probability distributions, thereby aiding in model assessment, feature selection, and optimization. In multiclass classification, where the goal is to predict the class label or probability distribution for each sample, KL divergence serves as a crucial measure for evaluating the dissimilarity between the predicted and true class distributions. By comparing the predicted probabilities with the actual class distribution, KL divergence provides insights into the model's performance, highlighting areas where the predictions deviate from the ground truth. Moreover, KL divergence can be leveraged for feature selection, helping identify informative features that contribute significantly to the predictive performance of the model. Additionally, KL divergence plays a vital role in optimization tasks, guiding the fine-tuning of model parameters to minimize the discrepancy between predicted and actual class distributions, thereby enhancing classification accuracy. Overall, KL divergence serves as a valuable tool in multiclass classification, facilitating model evaluation, feature selection, and optimization, ultimately leading to improved classification performance and robustness in real-world applications. The proposed PBGM model begins with the training using the traditional GBM model and the objective assignment is done with multiple classes as the problem deals with multiclass classification. As a probabilistic model, the outcomes of the GBM were analyzed for the divergence, and the predictive analysis was made which is unique when compared to the existing models.

4. RESULTS AND DISCUSSION

To analyze the working of the existing and proposed multiclass classification model we used the preprocessed water quality management dataset that consists of 13 features and 3,277 samples. Accuracy conveys the true prediction from the overall prediction that is done by the model. Precision is used to calculate the predictions that are made rightly by the model whereas recall is used to measure the percentage of data points that are related to the model that have been accurately identified. The F1 score is used to calculate the harmonic mean from both precision and recall to analyze the performance of the

classification model. The LGM model is a traditional model that is used for classification and prediction problems. In our research, the GBM model is used for training the model where it has produced an accuracy of 83.26% whereas the model has achieved a precision of 87%, recall of 85.57%, and f1 score was measured as 86.31%, respectively. The most important aspect of this classification could be the splitting between the multiple classes for prediction. As per the classification made by the model, 218 samples are in the Excellent category, 229 samples belong to Good category, and 233 samples belong to Poor category of water quality level observed from the feature vector. The initial training from the samples to the model has given a significant performance where there is a need for improvement in terms of accuracy. One more aspect of the analysis is to validate whether the model is prone to overfitting. To perform a better analysis, another linear model LGBM was executed for the feature vector which improved the accuracy of the model to 90.21%, the precision of the model reached 93.12%, the recall was 91.52%, and the f1 score was measured as 95.61%. The confusion matrix obtained by the prediction model is given in Figure 3 which shows the split between the three major classes used for classification. This improvement in performance has led to an analysis using an ensemble classification before which there must be some improvement in the training phase that has to be carried out using the boosting model. The XGBoost algorithm was finally implemented as the individual model for prediction from the obtained feature vector. This model has produced an accuracy of 95.62%, precision of 93.14%, recall of 92.71%, and the f1 score was measured as 93.92%. This model has a split between three major classes in water quality assessment data with 216 samples in the Excellent category, 229 samples in the Good category, and 230 samples reported in the Poor category. The training done using the existing models improved the overall performance of the multiclass classification concerning three different classes. The model training, testing, and validation with remaining samples is most essential for the analysis where the training dataset has 2,597 samples, and the remaining 680 samples are used for validation and testing. The existing and the proposed algorithms were executed using Python version 3.12.1 with the system specification of Intel(R) Core(TM) i7-8565U CPU; 12 GB RAM; Windows 10.

The overall performance of the water quality assessment was improved by constructing the proposed probabilistic gradient boost model (PGBM). This model had done the ensemble of LGBM classifier for model training whereas, in the testing and validation phase, the XGBoost along with probabilistic KLD method was adopted for the testing and validation of the

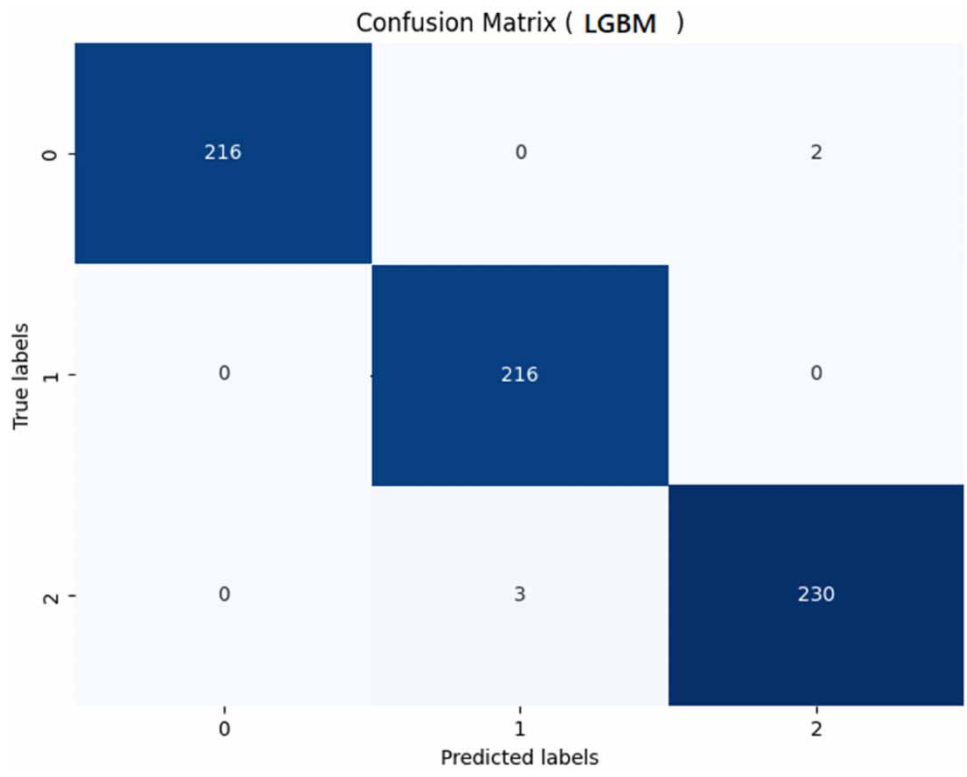


Figure 3 | Confusion matrix for LGBM prediction model.

proposed model. The prediction obtained through the proposed PGBM model has recorded the highest accuracy and precision of 98%, whereas the recall and f1 score of the proposed model recorded 99% which was the highest performance metrics that are achieved by any of the existing models that are used for multiclass classification of water quality level assessment so far. The confusion matrix obtained by the proposed PGBM model is shown in Figure 4, which accommodated two more testing samples in the low-level samples when compared to the traditional LGBM classifier.

From Figure 4, it is evitable that the class discrimination during the testing and validation phase of the proposed PGBM model is different when compared to the confusion matrix of the traditional LGBM prediction model. The comparison of the proposed PGBM model with the existing models that are used in recent research in water quality assessment [8,9] was also performed in terms of accuracy, specificity, and sensitivity. The comparison results are given in Table 1, which precisely conveys the models that are used in water quality assessment so far.

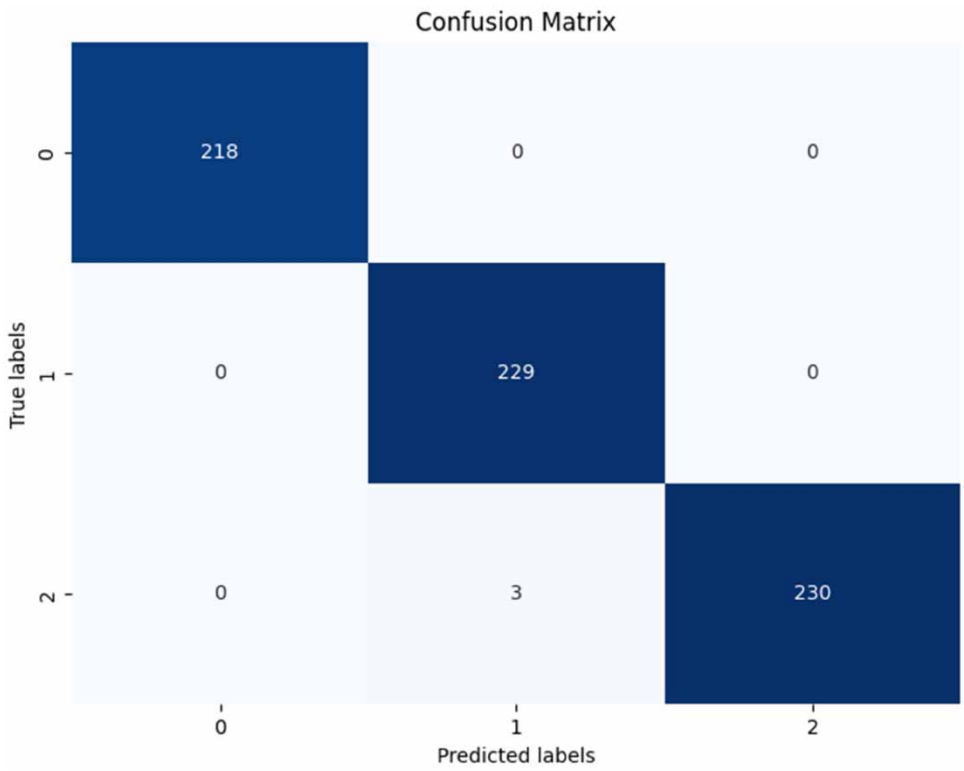


Figure 4 | Confusion matrix for the proposed PGBM classification model.

Table 1 | Performance metrics of the proposed and existing models

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
XGBoost	73.5	66.0	76.9
AdaBoost	74.4	64.0	76.9
SVM	75.5	60.9	79.0
K-NN	75.9	45.7	83.1
DT	78.3	80.3	79.5
LGBM	90.21	89.03	91.54
PGBM (proposed)	98	98	96

So far in the water quality assessment, six different models have been adopted and their performance was measured. XGBoost, AdaBoost, SVM, K-NN, DT, and LGBM have been implemented for the analysis of water quality assessment where LGBM has provided the highest performance in terms of accuracy of 90.21%, sensitivity of 89.03%, and specificity of 91.54%, respectively. Even though this has the highest performance, still 90.21% accuracy is not a considerable performance metric for quality assessment which in turn could be difficult for real-world implementation. The computation of three

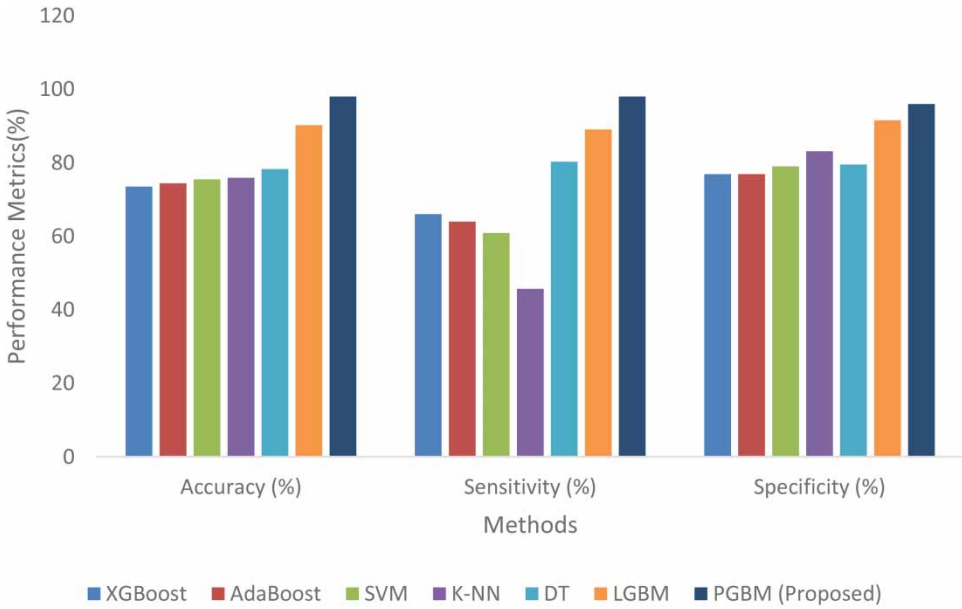


Figure 5 | Graphical representation of the proposed and existing models.

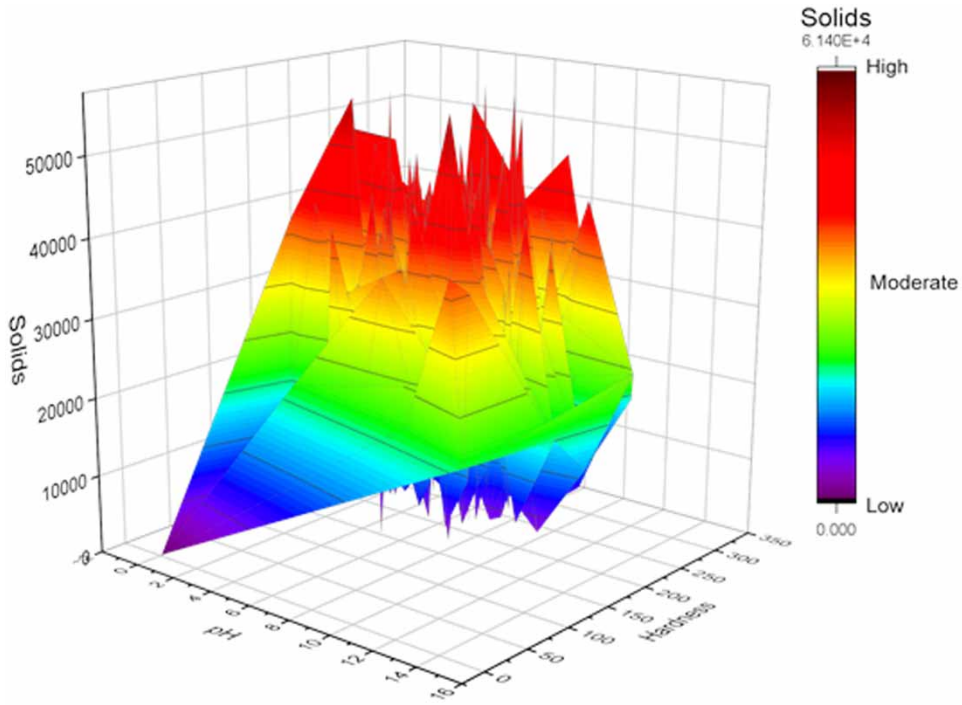


Figure 6 | 3D representation of similarity correlation among the features.

major features and incorporating it with the existing feature vector provided a performance improvement when it was implemented with the LGBM prediction model. LGBM was improved by 10.5% in accuracy, a sensitivity of 7.4%, and a specificity of 8.1% when compared to its earlier implementation without three computed features. Furthermore, performance improvement is ensured by the proposed probabilistic ensemble linear model (PGBM) prediction. The graphical representation of the proposed and existing models is given in Figure 5.

This model has achieved a remarkable accuracy of 98% due to the probabilistic voting mechanism by Kullback Libeler Divergence Method where the change in distribution between two divergences has led to an accurate classification. The model also achieved 98% sensitivity and 96% specificity. The major factor for performance improvement is due to the feature computation and correlation between the available features.

Figure 6 denotes the representation of available features along with their correlation for analyzing the obtained data. Three major features solids in the water, pH level of the water, and the hardness of the water are considered for combinatory analysis to calculate their impact in deriving the target vector. It can be noted that wherever the solids and hardness of the water are high the pH level of the water gets lower. Thus the proposed model is the highest-performance prediction model for water quality assessment with the proposed feature computation and extraction strategies.

5. CONCLUSION

Estimating the water quality levels by inducing three major features and initializing three major levels was the major focus of the article. We proposed an IoT framework to perform efficient data collection. The water quality assessment done so far analyzed the environmental aspects of the features whereas the proposed models calculated the statistical relationship between the features and the new set of features are computed using the proposed triple-stage feature computation mechanism. A multiclass classification model was proposed to classify the available samples into several classes that are extracted in the feature computation stage. We proposed a PGBM which outperformed all the other existing models in terms of accuracy, specificity, and sensitivity. Time series-based feature extraction and the implementation of deep learning models for water quality assessment levels in global level data will be our future research direction.

ACKNOWLEDGEMENTS

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was supported by the Deanship of Scientific Research, Vice President for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [GrantA377].

FUNDING

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (NURSP2024R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M. & Elshafie, A. 2019 Machine learning methods for better water quality prediction. *Journal of Hydrology* **578**, 124084.

Akbarighatar, P., Pappas, I. & Vassilakopoulou, P. 2023 A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review. *International Journal of Information Management Data Insights* **3** (2), 100193.

Aslam, B., Maqsoom, A., Cheema, A. H., Ullah, F., Alharbi, A. & Imran, M. 2022 Water quality management using hybrid machine learning and data mining algorithms: An indexing approach. *IEEE Access* **10**, 119692–119705.

Baek, S. S., Pyo, J. & Chun, J. A. 2020 Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water* **12** (12), 3399.

Bedi, S., Samal, A., Ray, C. & Snow, D. 2020 Comparative evaluation of machine learning models for groundwater quality assessment. *Environmental Monitoring and Assessment* **192**, 1–23.

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H. & Kazakis, N. 2020 Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment* **721**, 137612.

Cao, S., You, R., Li, X., Jia, J., Wang, J. & Liu, Y. 2022 A novel approach for estimating the capacity of ungauged small reservoirs using remote sensing and DEM. *Hydrology Research* **53** (7), 1001–1016.

Cerda, P., Varoquaux, G. & Kégl, B. 2018 Similarity encoding for learning with dirty categorical variables. *Machine Learning* **107** (8), 1477–1494.

Chaudhari, K. G. 2019 Water quality monitoring system using internet of things and swqm framework. *International Journal of Innovative Research in Computer and Communication Engineering* **7** (9), 3898–3903.

Chen, J., Yang, M., Gong, W. & Yu, Y. 2022 Multi-neighborhood guided Kendall rank correlation coefficient for feature matching. *IEEE Transactions on Multimedia* **25**, 7113–7127.

Condon, L. E., Kollet, S., Bierkens, M. F., Fogg, G. E., Maxwell, R. M., Hill, M. C., Fransen, H. J. H., Verhoef, A., Van Loon, A. F., Sulis, M. & Abesser, C. 2021 Global groundwater modeling and monitoring: Opportunities and challenges. *Water Resources Research* **57** (12), e2020WR029500.

Dai, H., Ju, J., Gui, D., Zhu, Y., Ye, M., Cui, J. & Hu, B. X. 2024a A two-step Bayesian network-based process sensitivity analysis for complex nitrogen reactive transport modeling. *Journal of Hydrology* **632**, 130903. https://doi.org/10.1016/j.jhydrol.2024.130903.

Dai, H., Liu, Y., Guadagnini, A., Yuan, S., Yang, J. & Ye, M. 2024b Comparative assessment of two global sensitivity approaches considering model and parameter uncertainty. *Water Resources Research* **60** (2), e2023WR036096. https://doi.org/10.1029/2023WR036096.

Dilmi, S. & Ladjal, M. 2021 A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques. *Chemometrics and Intelligent Laboratory Systems* **214**, 104329.

Gaur, N., Sarkar, A., Dutta, D., Gogoi, B. J., Dubey, R. & Dwivedi, S. K. 2022 Evaluation of water quality index and geochemical characteristics of surface water from Tawang India. *Scientific Reports* **12** (1), 11698.

Hosseini Baghanam, A., Norouzi, E. & Nourani, V. 2022 Wavelet-based predictor screening for statistical downscaling of precipitation and temperature using the artificial neural network method. *Hydrology Research* **53** (3), 385–406.

Jayaraman, P., Nagarajan, K. K., Partheeban, P. & Krishnamurthy, V. 2024 Critical review on water quality analysis using IoT and machine learning models. *International Journal of Information Management Data Insights* **4** (1), 100210.

Khan, M. S. I., Islam, N., Uddin, J., Islam, S. & Nasir, M. K. 2022 Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences* **34** (8), 4773–4781.

Konde, S. & Deosarkar, D. S. 2020 IOT based water quality monitoring system. In: *2nd International Conference on Communication & Information Processing (ICCIP)*.

Kruse, P. 2018 Review on water quality sensors. *Journal of Physics D: Applied Physics* **51** (20), 203002.

Kumar, M. & Padhy, P. K. 2014 Multivariate statistical techniques and water quality assessment: Discourse and review on some analytical models. *International Journal of Environmental Sciences* **5** (3), 607–626.

Liu, S., Wang, J., Wang, H. & Wu, Y. 2022 Post-processing of hydrological model simulations using the convolutional neural network and support vector regression. *Hydrology Research* **53** (4), 605–621.

Liu, Y., Fang, Z., Cheung, M. H., Cai, W. & Huang, J. 2023 Mechanism design for blockchain storage sustainability. *IEEE Communications Magazine* **61** (8), 102–107. doi:10.1109/MCOM.001.2200809.

Luo, J., Zhao, C., Chen, Q. & Li, G. 2022 Using deep belief network to construct the agricultural information system based on Internet of Things. *The Journal of Supercomputing* **78** (1), 379–405. doi: 10.1007/s11227-021-03898-y.

Marín Celestino, A. E., Martínez Cruz, D. A., Otazo Sánchez, E. M., Gavi Reyes, F. & Vásquez Soto, D. 2018 Groundwater quality assessment: An improved approach to K-means clustering, principal component analysis and spatial analysis: A case study. *Water* **10** (4), 437.

Monica, N. & Choi, K. 2016 Temporal and spatial analysis of water quality in Saemangeum watershed using multivariate statistical techniques. *Paddy and Water Environment* **14**, 3–17.

Najwa Mohd Rizal, N., Hayder, G., Mnzool, M., Elnaim, B. M., Mohammed, A. O. Y. & Khayyat, M. M. 2022 Comparison between regression models, support vector machine (SVM), and artificial neural network (ANN) in river water quality prediction. *Processes* **10** (8), 1652.

Nayak, J. G., Patil, L. G. & Patki, V. K. 2020 Development of water quality index for Godavari River (India) based on fuzzy inference system. *Groundwater for Sustainable Development* **10**, 100350.

Pandey, G., Ren, Z., Wang, S., Veijalainen, J. & de Rijke, M. 2018 Linear feature extraction for ranking. *Information Retrieval Journal* **21** (6), 481–506.

Qi, Z., Wang, G. & Zhang, B. 2022 Study on the transformation of surface water and groundwater in the water source area of Baima-Jili River Basin. *Hydrology Research* **53** (4), 622–637.

Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., Maalouf, S. & Adams, C. 2020 Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews* **205**, 103187.

Tang, X., Wang, Z., He, Q., Liu, J. & Ying, Z. 2020 Latent feature extraction for process data via multidimensional scaling. *Psychometrika* **85** (2), 378–397.

Thorslund, J. & van Vliet, M. T. 2020 A global dataset of surface water and groundwater salinity measurements from 1980–2019. *Scientific Data* **7** (1), 231.

Uddin, M. G., Nash, S. & Olbert, A. I. 2021 A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators* **122**, 107218.

Uddin, M. G., Nash, S., Rahman, A. & Olbert, A. I. 2023 Assessing optimization techniques for improving water quality model. *Journal of Cleaner Production* **385**, 135671.

Wang, X., Zhang, F. & Ding, J. 2017 Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific Reports* **7** (1), 12858.

Wu, Z., Zhang, D., Cai, Y., Wang, X., Zhang, L. & Chen, Y. 2017 Water quality assessment based on the water quality index method in Lake Poyang: The largest freshwater lake in China. *Scientific Reports* **7** (1), 17999.

Wu, Z., Lai, X. & Li, K. 2021 Water quality assessment of rivers in Lake Chaohu Basin (China) using water quality index. *Ecological Indicators* **121**, 107021.

Zhang, K., Li, Y., Yu, Z., Yang, T., Xu, J., Chao, L., Ni, J., Wang, L., Gao, Y., Hu, Y. & Lin, Z. 2021 Xin'anjiang nested experimental watershed (XAJ-NEW) for understanding multiscale water cycle: Scientific objectives and experimental design. *Engineering* **18** (11). doi:10.1016/j.eng.2021.08.026.

Zhou, D., Sheng, M., Bao, C., Hao, Q., Ji, S. & Li, J. 2024 6G non-terrestrial networks-enhanced IoT service coverage: Injecting new vitality into ecological surveillance. *IEEE Network*. doi:10.1109/MNET.2024.3382246.

First received 9 April 2024; accepted in revised form 31 May 2024. Available online 4 July 2024