Earth Science Informatics (2023) 16:1701–1725
https://doi.org/10.1007/s12145-023-00977-x

RESEARCH

Groundwater Quality Analysis and Drinkability Prediction using Artificial Intelligence

Niranjan Panigrahi[1] · S. Gopal Krishna Patro[2] · Raghvendra Kumar[3] · Michael Omar[4] · Tran Thi Ngan[5] · Nguyen Long Giang[6] · Bui Thi Thu[6] · Nguyen Truong Thang[6]

Received: 10 June 2022 / Accepted: 18 February 2023 / Published online: 4 April 2023

Abstract

Water quality strongly influences sustainable growth of a healthy society and green environment. According to the International Initiative on Water Quality (IIWQ) of the UNESCO Intergovernmental Hydrological Programme (IHP), it is essential to address water-quality issues holistically in developed and developing countries. Due to rapid urbanization and industrialization in many developing countries, groundwater - one of the major sources of drinking is getting highly affected. The traditional laboratory-based chemical testing process with conventional statistical methods is often used to analyze water quality. However, it is time-consuming. Recently, Artificial Intelligence (AI) based approaches have proven to be a better alternative for analysis and prediction of the quality of water, provided with its chemical components' data. In this paper, we present research focusing on groundwater quality analysis using Artificial Intelligence (AI) in a case study of Odisha, an eastern- state of India and the data acquired from the Northern delta, the North Central Coast of Vietnam. The dataset in Vietnam is collected by the Ministry of Natural Resources and Environment, providing technical regulations on water resources monitoring. The Central Groundwater Board and the Government of India collect the dataset from India. The target problem is formulated as a multi-class classification problem to predict groundwater quality for drinking suitability by WHO standards. AI methodologies such as logistic regression, K-NN, Support Vector Machine (SVM) variants, decision tree, AdaBoost and XGBoost are used. Prediction results have demonstrated that Adaboost, the XGBoost and the Polynomial SVM model accurately classified the Water Quality Classes with an accuracy of 92% and 98%, respectively. It would help decision-makers effectively choose the best source of water for drinking.

Keywords Water quality analysis · Drinkability prediction · Machine learning · AdaBoost · SVM · XGBoost

Introduction

With growing urbanization and deforestation, groundwater quality is changing invisibly due to various types of pollution. It is because of harmful substances like chemicals, microorganisms, radio-activity or heat energy which enter directly or indirectly into bodies of water. Classifying and predicting water quality is important for various purposes like drinking, and irrigation (Wang et al. 2017; WHO Guidelines for drinking water quality 2004). Recently, interdisciplinary research has gained momentum to study groundwater quality across various parts of the globe (Venkata Vara Prasad et al. 2022; Dogo et al. 2019; Ranjithkumar and Robert 2021). The major focus has been given to developing countries due to the fact that industrialization and urbanization in developing countries are mostly affecting groundwater quality (Ground Water Year Book 2018).

Water quality is usually assessed by costly, traditional laboratory and statistical analysis, which is time-consuming (WHO Guidelines for drinking water quality 2004). Several types of research (Sahu et al. 2021; Barik and Pattanayak 2019; Ground Water Year Book 2018; Madhav et al. 2020) have been done to carry out water quality analysis focusing on only hydro-chemical processes. In this regard, AI-based approaches can be used for quick and reliable analysis and have been recently adopted by research communities across the globe (Hanoon et al. 2021; Ahmed et al. 2020; Khan and See 2016). This solves major issues related to water.

Communicated by: H. Babaie

✉ Nguyen Long Giang
nlgiang@ioit.ac.vn

Extended author information available on the last page of the article.

A set of state-of-the-art ML models are selected, which have shown better performance on many regional water quality datasets of different countries. Motivated by their performance, the present work has considered applying those models and their variants and studying the efficacy of the models on the Odisha, India and Vietnam water quality datasets.

In Sahu et al. (2021), processes using silicate-halite dissolution and reverse ion exchange were employed for the phreatic aquifer, considering Odisha as a focused area. In Barik and Pattanayak (2019), the authors used data plot dispositions on Gibb's diagram to indicate the chemistry of groundwater for irrigation purposes in Rourkela city of Odisha. In Harichandan et al. (2021), empirical correlation analysis between WQI and physio-chemical parameters was investigated to study the drinkability of water. The authors in (Madhav et al. 2020) applied hydro-chemical processes to study drinking and farming cases.

To the best of our knowledge, machine learning based approaches have so far not been incorporated for water quality analysis, especially in Odisha and Vietnam, which can make the process more effective and less time consuming. In this context, some machine learning approaches that have already been applied successfully in other regions of the world are presented below as a point of motivation.

In Wang et al. (2021), stream water quality was predicted for different urban densities scenario using explainable machine learning methods. The authors in Haghiabi et al. (2018) have predicted WQI by random forest method, namely ANN and SVM. A similar approach was presented in Kouadri et al. (2021) on irregular datasets for the southeast Algerian region using multi-linear regression, random forest, M5P tree etc. The WQI based ML methods were also used in Wang et al. (2017) for the Ebinur lake watershed, China.

Supervised learning methods were used in Ahmed et al. (2019) for water-quality analysis of Rawal water lake in Pakistan. In Theyazn et al. (2020), the authors used AI based approach using an auto-regressive neural network model named NARNET for water quality analysis and classification. Principal component analysis (PCA) and gradient boosting methods were used in Khan et al. (2021) for water quality prediction and classification. Using hydro-meteorological data, a data-driven model was proposed in Sokolova et al. (2022) for predicting microbial water quality.

In Tiyasha et al. (2021), the authors have focused on assessing Klang river water quality using deep learning models. It involves water quality index computation by considering six notable water quality parameters. The proposed method uses random forest, decision tree and deep learning models on two scenarios: "small scale catchment" and "large scale catchment". In both cases, the deep learning model is claimed to perform well in the case of non-linear data.

The authors in Tiyasha et al. (2020) have presented a thorough survey for the last decade on AI based model development for river water quality assessment. The major points focused on in this survey are variability in inputs for river water quality assessment, model architecture, and metrics for evaluation and investigation in different regions.

In Tiyasha et al. (2021), the authors adopted a hybrid tree-based approach for predicting river dissolved oxygen (DO) using satellite and hydro-meteorological data for the Klang river of Malaysia. Different selector algorithms are used for feature selection, namely, Boruta, GA, MARS and XGBoost. In the next phase, tree-based models like random forest, Ranger, and cForest are used to predict the DO. The best-performing models reported are XGBoost and MARS while considering the coefficient of determination as the evaluating parameter.

In the study by Nizal et al. in Nur Najwa Mohd et al. (2022), water quality parameters are predicted using a neural network-based approach with an integrated GUI. The focused region is selected as the Langat River of Malaysia. They adopted a novel approach of including rainfall data to predict water quality. The GUI design takes real-time inputs and can predict different water quality parameters.

Ubah et al. (Ubah et al. 2021) have proposed ANN-based models for analyzing river water quality for irrigation purposes. The data are collected for Ele river Nnewi of Anambra State. The model is capable of predicting the water quality index for one year. The authors in Venkata Vara Prasad et al. (2021) propose an automated analysis of water quality using ML and autoML methodology. They claimed that autoML performs better than conventional ML if binary classes are predicted. The authors in Zhu et al. (2022) have presented an extensive survey on water quality analysis for different environments like drinking water, surface water, seawater etc. A set of 45 ML algorithms are evaluated and presented for water quality analysis.

In this paper, we make a reasonable attempt to use AI-learning-based models to analyse the water quality for drinking purposes.

We applied XGboost, a polynomial support vector machine, a decision tree, logistic regression, a K-NN, and a CNN in this experiment. It has been found that XGBoost performs best, with an accuracy of 92.67 and 98%; as mentioned in the paper, we have provided more detailed results of the other models in Section "Results summary & discussion", Results summary and Discussion.

The dataset collected and published by the Central Ground Water Board, Government of India and the Ministry of Natural Resources and Environment providing technical regulations on water resources monitoring in Vietnam are considered as inputs for this case study (Ground Water Year Book 2018). The significant contributions of this research are depicted below.

i. The underlying problem is formulated as a multi-class classification problem for a distinct classification of the drinkability of water
ii. State-of-the-art water quality estimation models, including Water Quality Index (WQI) model and Water Quality Class (WQC) model as per WHO specifications, are used to carry out a realistic analysis
iii. A thorough exploratory data analysis is carried out for better water quality prediction
iv. A set of well-known learning models are used for optimal prediction of WQC

The rest of the paper follows: Section "Water quality estimation model" describes the water quality estimation model, and Section "Problem formulation and proposed framework" shows the problem formulation and proposed framework. Section "Proposed strategy" explains the detailed strategy adapted to apply the AI-learning-based approaches. Section "Results summary & discussion" presents the results summary with performance evaluation metrics; the conclusion is given in Section "Conclusion".

Water quality estimation model

The water quality index (WQI) is used to measure the quality of water, and it is calculated based on some state-of-the-art parameters (Khan and See 2016; Haghiabi et al. 2018; Wang et al. 2017). To estimate WQI, mostly nine well-known parameters are considered. In our case, out of four- teen surveyed parameters as given in Ground Water Year Book (2018), after performing exploratory data analysis, the thirteen most influencing parameters are considered, including Total Hardness (TH), Total Dissolved Solids (TDS), pH, Sulphate (SO_4), Electrical Conductivity (EC), Alkalinity, Magnesium (Mg), Sodium (Na), Potassium (K), Chloride (Cl), Calcium (Ca), Fluoride (F), and Bicarbonate (HCO_3). As per WHO guidelines (WHO Guidelines for drinking water quality 2004), the permissible range of different parameters is shown in Table 1. In the case of the Vietnam dataset, out of twenty-one surveyed parameters as given by the Vietnam authorities, after performing exploratory data analysis, 12 most influencing parameters are considered, including Total Dissolved Solids (TDS), pH, Sulphate (SO_4), Harshness-General, Harshness-Permanent, Magnesium (Mg), Sodium (Na), Potassium (K), Chloride (Cl), Calcium (Ca), Fluoride (F), Bicarbonate (HCO_3). Table 2 shows the permissible range of different parameters. Using these parameters and prescribed weights, the WQI and WQC are defined for each sample as given below.

Table 1 Odisha Permissible value for parameters used in calculating WQI

Parameter	Permissible Value
PH	6.5-8.5
Sodium	200
Electrical Conductivity(EC)	1000
Potassium	12
Total Dissolved Solids(TDS)	500
Carbonate	0
Total Hardness	300
Bicarbonate	350
Alkalinity	200
Chloride	250
Calcium	75
Sulphate	200
Magnesium	50
Fluoride	1.5

Water quality index model

To estimate the water quality, a standard parameter named as Water Quality Index (WQI) given by the equation (1) is mostly used (Kouadri et al. 2021) (Haghiabi et al. 2018). In this study, a total of 13 from Odisha and sixteen from

Table 2 Vietnam Permissible value for parameters used in calculating WQI

Parameter	Permissible Value
PH	5.5 - 8.5
$Sodiumna$	200
$Fe3iron$	5
$Potassiumk$	12
$TotalDissolvedSolids$	400
$Carbondioxideco2_free$	51
$SulphateironSo4$	400
$Bicarbonate$	350
$Harness_general$	300
$Chlorinecl$	250
$Calcium$	75
$Sulphate$	200
$Magnesiummg2$	50
$Fluoride$	1
$Harness_permanent$	393
$TotalPhenol$	0.001

Vietnam significant parameters are taken into consideration to calculate the WQI.

$$WQI = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i=1}^{N} w_i} \tag{1}$$

where N represents the total number of parameters considered for water quality evaluation (In our case, it is 13 in Odisha and 12 in Vietnam). q_i is the quality rating scale for the individual parameters, which is computed using equation (2).

$$q_i = 100 \times (\frac{V_i - V_{ideal}}{S_i - V_{ideal}}) \tag{2}$$

In the above equation, V_i is the estimated value for parameter 'i' and V_{ideal} is the permissible value for parameter 'i' in case of water without impurity. S_i is the permissible value for parameter 'i'. Further, w_i is estimated as given in equation (3).

$$w_i = \frac{K}{S_i} \tag{3}$$

where K is a proportionality constant which is calculated below.

$$K = \frac{1}{\sum_{i=1}^{N} S_i} \tag{4}$$

Water quality class model

Using WQI values, WHO (WHO Guidelines for drinking water quality 2004) has prescribed a standard range for Water Quality Class (WQC) of drinking water which is given below.

$$WQC = \begin{cases} Excellent, & if\ 0 \leq WQI \leq 25, \\ Good, & if\ 26 \leq WQI \leq 50, \\ Poor, & if\ 51 \leq WQI \leq 75, \\ Very\ Poor, & if\ 76 \leq WQI \leq 100, \\ Undrinkable, & if\ WQI > 100 \end{cases} \tag{5}$$

Problem formulation and proposed framework

This section briefly presents the problem formulation for water quality analysis and drinkability prediction using the above water quality estimation models. After that, a framework is proposed to solve the problem using machine learning based approaches.

Water drinkability as Multi-class classification problem

Based on the WQI value which can be estimated as in equation (1) and mapping it to a respective class in equation (5), the water quality for drinking purposes can be predicted. Indeed, using equation (5), it is formulated as a classification problem with multiple classes as per the range of WQI value. Assuming the training dataset consists of 'n' features or attributes of water quality, formally, the problem can be expressed as:

Each data sample is represented as a pair of $\{(x_i, Y_i)\}_{i=1}^{n}$, where x_i are vectors of features and $Y_i \in \{1, 2, .., k\}$, the respective labels representing one of the k classes according to estimated WQI value. In our research, for more meaningful classification in terms of drinkability, the water quality classes are modified as given in the equation below without loss of generality. The modified water quality class model used in our work is given as in equation (6) with $k = 4$ and $k = 10$, including Excellent, Good, Poor, Bad classes for the Odisha dataset while Excellent, Good, Medium, Poor, Fair for the Vietnamese dataset.

$$OdishaWQC = \begin{cases} Excellent, & if\ 0 \leq WQI \leq 25, \\ Good, & if\ 26 \leq WQI \leq 50, \\ Poor, & if\ 51 \leq WQI \leq 75, \\ Bad, & if\ WQI > 75 \end{cases} \tag{6}$$

$$VietnamWQC = \begin{cases} Poor, & if\ 0 \leq WQI \leq 25, \\ Fair, & if\ 25 \leq WQI \leq 50, \\ Medium, & if\ 50 \leq WQI \leq 70, \\ Good, & if\ 70 \leq WQI \leq 90, \\ Excellent, & if\ 90 \leq WQI \leq 100, \end{cases} \tag{7}$$

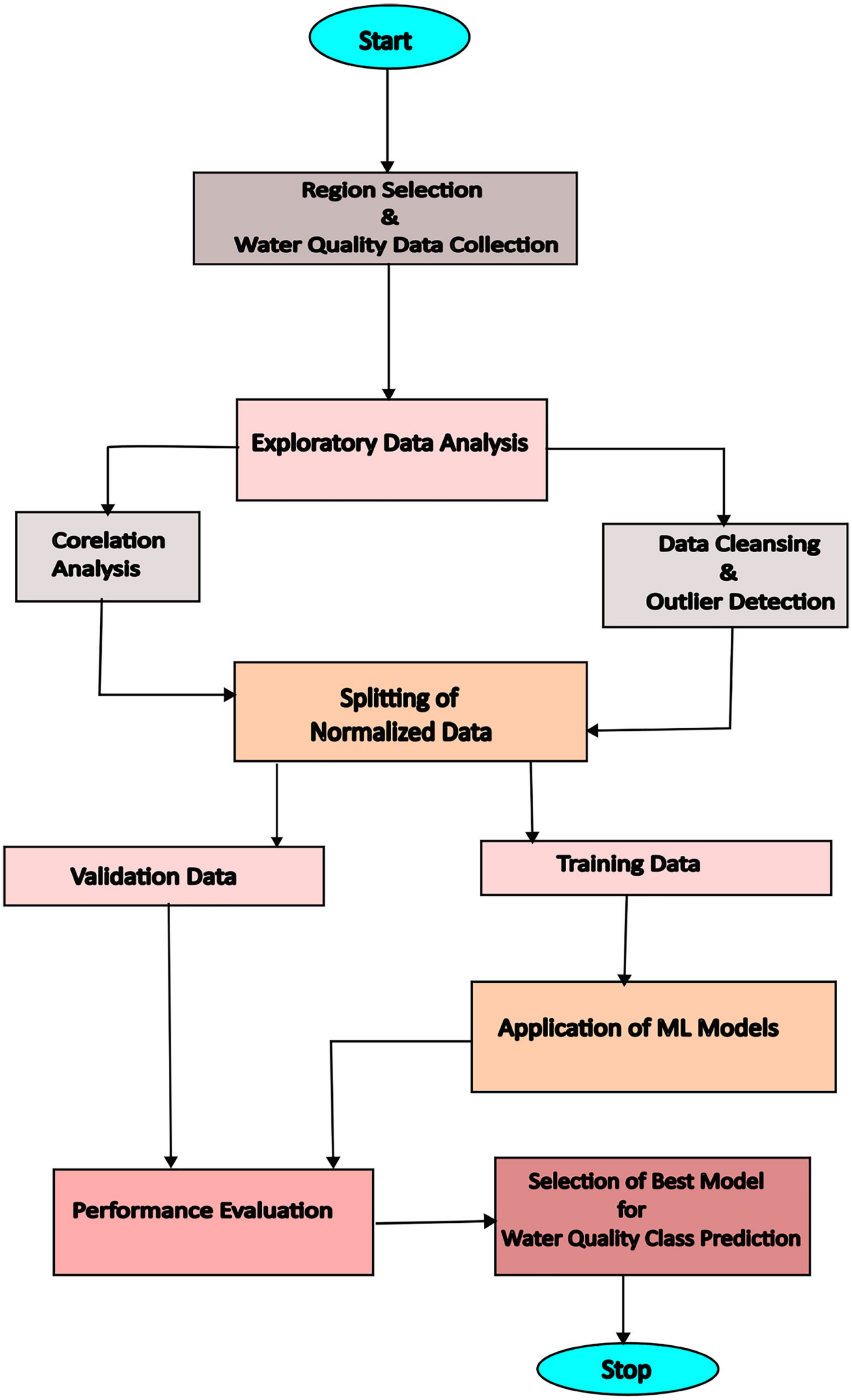
The proposed machine learning-based framework is depicted in the next section using the above description.

Proposed flow diagram

To solve the water quality analysis and drinkability problem, which is formulated as a multi-class classification problem as mentioned above, the schematic flow diagram of applying machine learning models is shown in Fig. 1.

It broadly goes through three phases: (i) Selection of the geographical region for which the groundwater quality needs to be analyzed and then the collection of water quality parameters value. (ii) In the second phase, Exploratory Data Analysis (EDA) and pre-processing of the data samples

Fig. 1 Proposed flow diagram for water quality analysis and prediction



are carried out. The EDA process mainly has two sub-steps: (a) correlation analysis, and (b) data cleansing & outlier detection. The correlation analysis involves determining the dependencies between various parameters

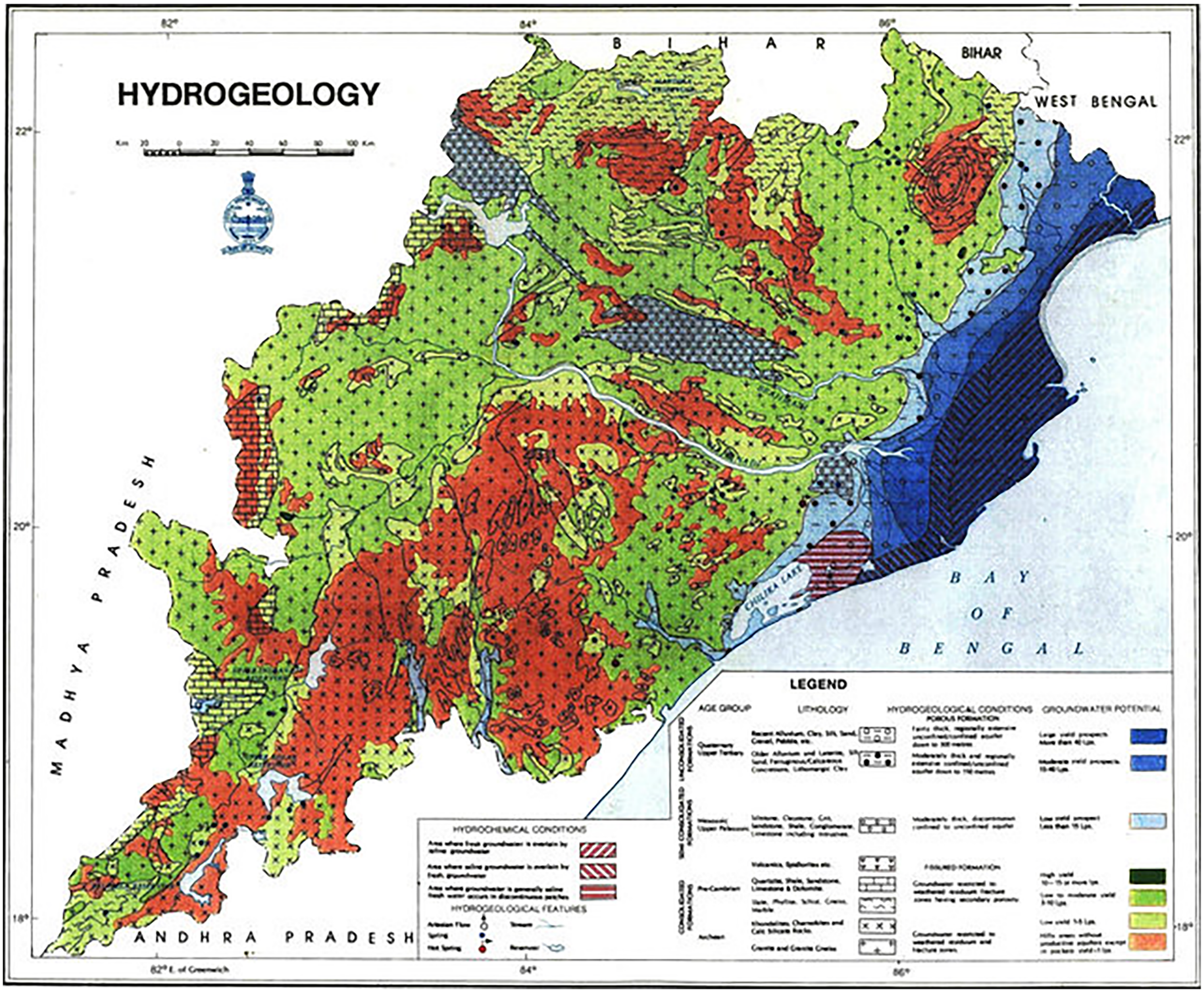


Fig. 2 Hydro-geological map of Odisha

and with WQI value. Subsequently, in the data cleansing & outlier detection phase, the non-contributing parameters are removed, and abnormal data values are ignored as outliers for better prediction. (iii) In this last phase, the data values are normalized for uniform scaling and then fed to applied ML models. The models are then evaluated using standard performance metrics. This evaluation will help decide the best model for predicting the water quality for the selected geographical region.

Proposed strategy

Following the framework shown in Fig. 1, this section describes the systematic process of predicting water quality with the specified drinkability classes using some AI-learning models.

Region selection and dataset collection

This research uses datasets from Odisha, an eastern state of India and the Northern delta and North central of Vietnam. The hydro-geological map of the state of Odisha is shown in Fig. 2. To monitor the ground-water level and chemical components, CGWB (Central Ground Water Board), Bhubaneswar, has set up 1600 NHNS (National Hydrograph Network Stations) as open/dug wells and piezometers as shown in Fig. 3. To observe the changes in the chemical components of groundwater, the water samples are collected once a year from these NHNS and analysed in the regional water laboratory. This data is further studied and represented in the Ground Water Year Book Report (Ground Water Year Book 2018). Thus, the data is extracted from the published report of 2018-19 and converted suitably to a CSV file for AI-based analysis. The dataset includes

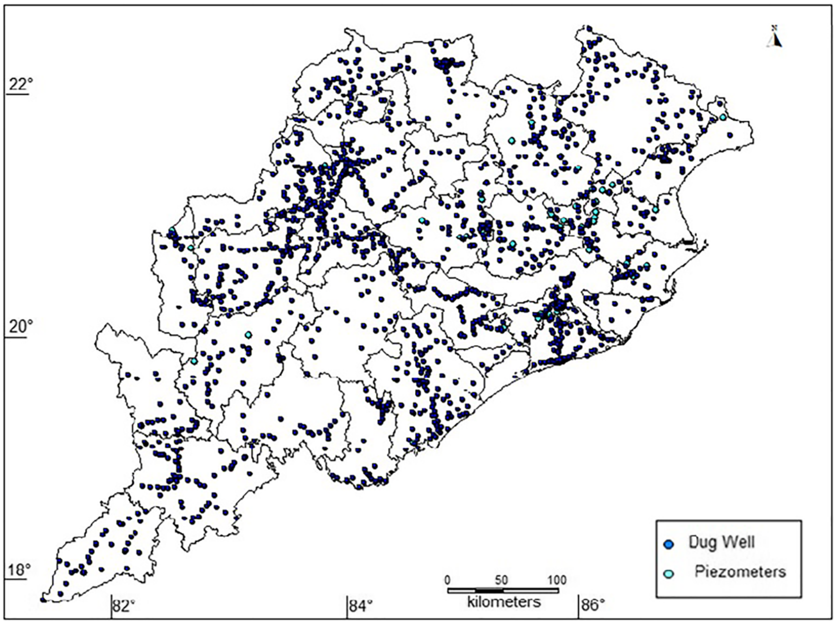


Fig. 3 National Hydrograph Network Stations in Odisha

13 parameters. In the case of Vietnam, 12 parameters are inclusive; thus, the data is extracted from the national water resources planning and inspection centre federation for planning and investigation of water resources in the North.

Exploratory data analysis

The Exploratory Data Analysis (EDA) is a pre-requisite process which is carried out to investigate data for patterns discovery, outlier detection, testing hypotheses and creating summary statistics and representing them graphically. The statistical description of the water sample dataset collected from Odisha is given in Tables 3, 4 and 5, and the statistical description of the water sample collected from Vietnam is given in Tables 6, 7 and 8 below.

The total number of samples collected is 1241 from all 30 districts of Odisha, while the samples collected in Vietnam are 2138. Table 3–8 shows some standard statistical metrics for the sample dataset, namely, the total number of samples (count), the mean value of a respective parameter, standard deviation (std), minimum (Min.) and maximum (Max.) parameter value and percentile-based description, considering 25%, 50 % and 75% of data samples for a particular parameter. This helps in getting an early insight into thenature of the samples. For example, it is observed from table 3 that the 50 percentile of the data has a pH value less than or equal to 7.9 which gives an inference that the quality is acidic and not good for drinking. Table 6–8 showing the parameters for the Vietnamese dataset, it is observed from Table 6 that the 50 percentile of the data has a pH value less than or equal to 7.0 giving a similar inference that the quality is acidic and not good for drinking.

Similarly, it is observed from Table 3 that the 50 percentile of the sample has EC less than or equal to 550. Also, it is inferred that the 50 percentile and 75 percentile of the samples have a Carbonate value of zero. So, this is considered a non-contributing parameter for water quality analysis and hence, not considered. The summary of all the metrics is presented by considering thirteen parameters, excluding Carbonate as the non-contributing parameter for

Table 3 Statistical description of groundwater dataset of Odisha

Metrics	PH	EC	TDS	TH	Alkalinity
Count	1241	1241	1241	1241	1241
Mean	7.82	695.4	358.057	215.015	178.496
Std.	.399	536.819	280.979	156.787	104.932
Min.	6.46	7.15	30	20	15
25%	7.58	360	186	123	105
50%	7.9	550	277	184	158
70%	8.12	900	456	267	228
Max.	8.78	5770	2766	1945	765

Table 4 Statistical description of the groundwater dataset of Odisha (continue)

Metrics	Ca	Mg	Na	Potassium
Count	1241	1241	1241	1241
Mean	43.941	25.617	49.991	13.162
Std.	30.612	25.835	61.033	29.670
Min.	0	-4	0	0
25%	26	10	17	1.6
50%	39	19	30	3.8
70%	53	34	65	10.1
Max.	497	345	820	332

Table 5 Statistical description of the groundwater dataset of Odisha (continue)

Metrics	Bi-carbonate	Chloride	Sulphate	Fluoride
Count	1241	1241	1241	1241
Mean	215.799	92.170	26.223	.398
Std.	126.507	127.14	30.662	.419
Min.	18	0	-3	0.02
25%	128	26	4	0.16
50%	192	55	17	0.27
70%	275	110	38	0.47
Max.	933	1753	434	3.94

Table 6 Statistical description of the groundwater dataset of Vietnam

Metrics	PH	K	TDS105	CO2-free	CL
Count	2138	2138	2138	2138	2138
Mean	6.974649	12.196974	1239.541553	51.047629	554.431787
Std.	0.680514	25.786524	2505.418750	63.327026	1448.927194
Min.	2.530000	0.080000	43.000000	0.000000	4.430000
25%	6.620000	2.100000	251.000000	17.585000	22.000000
50%	7.070000	4.700000	460.500000	35.200000	60.270000
70%	7.330000	10.000000	1110.250000	63.800000	449.332500
Max.	8.650000	305.000000	25901.000000	1573.000000	14623.130000

Table 7 Statistical description of the groundwater dataset of Vietnam (continue)

Metrics	Na	CO2-free	So4	Mg2
Count	2138	2138	2138	2138
Mean	319.902713	319.902713	33.693891	51.826906
Std.	824.533900	63.327026	114.156348	81.516255
Min.	0.800000	0.000000	0.000000	0.000000
25%	22.825000	17.585000	0.000000	12.162427
50%	52.600000	35.200000	4.800000	27.969214
70%	251.750000	63.800000	19.210000	55.933573
Max.	8200.000000	1573.000000	1218.000000	981.920000

Table 8 Statistical description of the groundwater dataset of Vietnam (continue)

Metrics	NO3-	HCO3-	HP(CaCO3)
Count	2138	2138	2138
Mean	0.3849119	353.486333	393.290115
Std.	1.750676	309.806085	679.036660
Min.	0.000000	0.000000	0.000000
25%	0.000000	140.350000	73.125000
50%	0.000000	271.540000	168.150000
70%	0.020000	475.960000	380.000000
Max.	44.800000	1742.120000	6445.000000

analysis. Further, since most of the Carbonate values are zero, any non-zero value found in the sample dataset is treated as an outlier.

Data cleansing and outlier detection in Odisha and Vietnam Dataset

The abnormal or outlier data points are removed by following a boxplot analysis. We set the max threshold as per the drinkability standards by the Government of India and standards set by the Vietnam authorities. Here we can observe in Fig. 4 using a boxplot that there are some readings of Fluoride and pH which exceed the general margin by a huge amount, in our case it is a natural occurrence that can cause some serious problems to the health of a person, a similar analysis was done for some other parameters as well while in Fig. 5, shows the number of outlier data point in Vietnam dataset are very few and not far from threshold values.

Correlation analysis in Odisha and Vietnam Dataset

Correlation analysis finds chemicals that depend on each other, which is needed to predict the existence of a chemical component if data is missing or changes suddenly. Tables 9 and 10 displays the corelation analysis parameters used in Odisha and Vietnam. We use Pearson Correlation. The degree of association or correlation and the type of relationship (Positive or Negative Correlation) help predict the presence/absence of interdependent chemicals. Figures 6 and 7 show correlation analysis results.

The following observations are made from the analysis Odisha dataset. (i) Calcium and Magnesium are correlated to TH (Total Hardness) - It is because TH is a measurement of calcium, magnesium and chloride. As hardness is caused due to chlorides of magnesium and calcium, our body needs both Ca and Mg to remain healthy.

(ii) Sodium and Magnesium are correlated to Chloride - From this, we can conclude that generally, in the water,

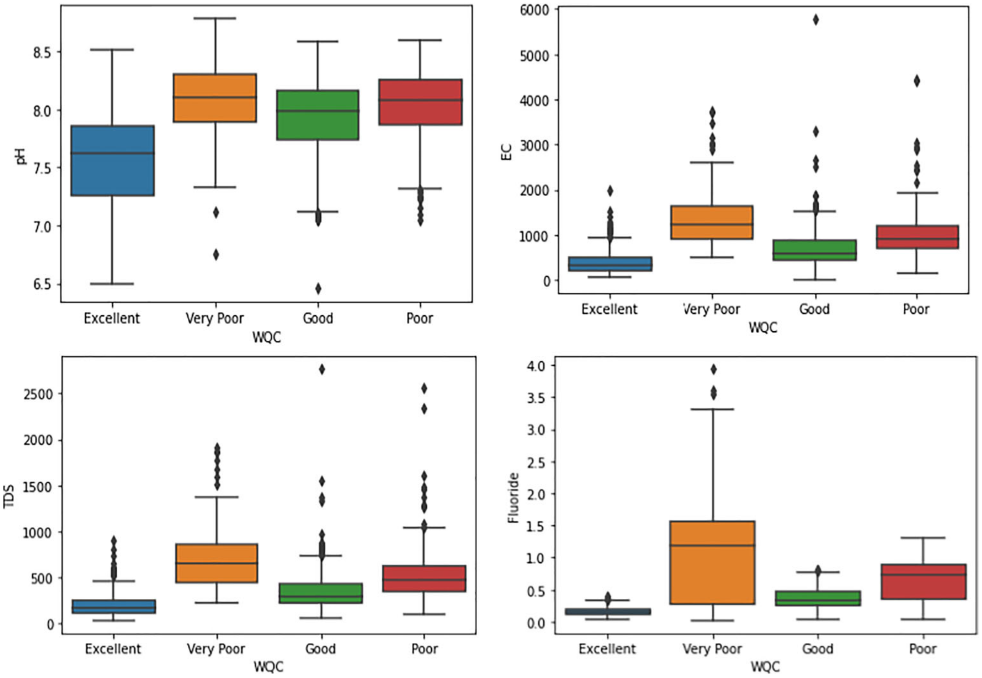


Fig. 4 Outlier Detection in the groundwater dataset Odisha

Fig. 5 Outlier Detection in the groundwater dataset Vietnam

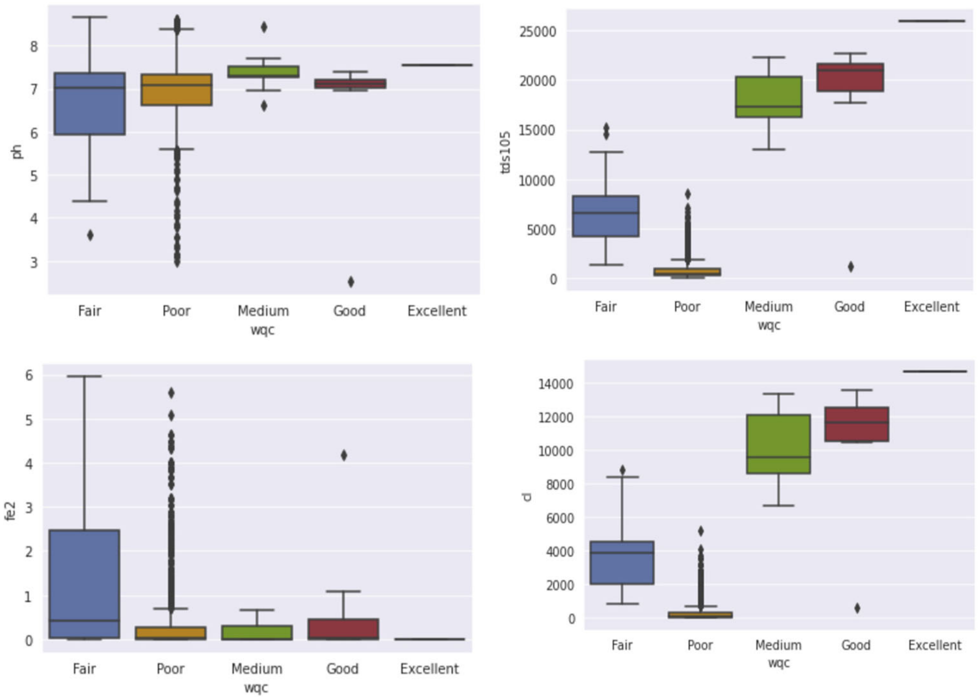


Table 9 Observation from co-relation analysis of different water parameters Odisha

Co-relation	pH	EC	TDS	TH	Alkalinity
pH	1.000000	0.151216	0.16141	0.107877	0.298009
EC	0.151216	1.000000	0.985947	0.85203	0.676578
TDS	0.16141	0.985947	1.000000	0.826257	0.676252
TH	0.107877	0.85203	0.826257	1.000000	0.512589
Alkalinity	0.298009	0.676578	0.676252	0.512589	1.000000
Calcium	0.049762	0.650262	0.641328	0.796832	0.335106
Magnesium	0.123624	0.789482	0.757618	0.901382	0.515389
is-drinkable.	-0.177312	-0.618706	-0.610951	-0.538213	-0.655714
WQI	0.40007	0.580675	0.600764	0.377814	0.640992
WQC	0.445067	0.568321	0.578188	0.418977	0.651174

Table 10 Observation from co-relation analysis of different water parameters Vietnam

Co-relation	Na	Mg2	Fe3	Fe2	Cl
Na	1.000000	0.926813	-0.065447	0.04619	0.99208
Mg2	0.926813	1.000000	-0.068776	0.061315	0.93432
Fe3	-0.065447	-0.068776	1.000000	0.065451	-0.05983
Fe2	0.04619	0.061315	0.065451	1.000000	0.077844
Cl	0.99208	0.93432	-0.05983	0.077844	1.000000
S04	0.331313	0.455828	0.02015	-0.008358	0.306298
pH	0.058535	0.056621	-0.153673	-0.493187	0.022139
is-drinkable.	-0.250264	-0.352268	0.000378	-0.088832	-0.253682
WQI	0.889289	0.929837	0.013497	0.291981	0.902264

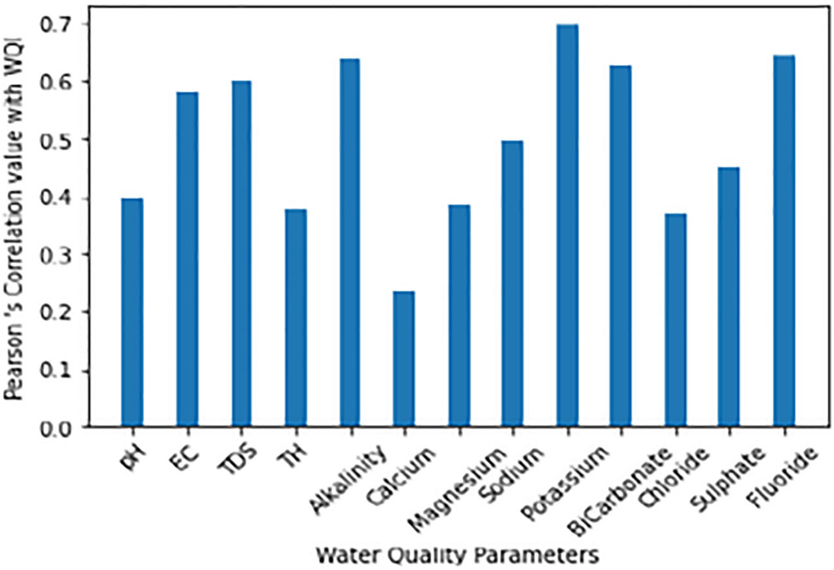


Fig. 6 Odisha co-relation analysis of WQI with all other water parameters

we have Sodium and Magnesium as Sodium Chloride and Magnesium Chloride. Both these salts are essential requirements for our body.

While the following observation was observed on the Vietnam dataset. There were strong correlations (a total of 17) between the pair of feature variables. After the bi-variate analysis, we dropped one of the features against each of the strong correlation matrices.

(i) Sodium and Magnesium are correlated to Chloride - From this, we can conclude that generally, in the water, we have Sodium and Magnesium as Sodium Chloride and Magnesium Chloride. Both these salts are essential requirements for our body Tables 16 and 17.

It is observed from Fig. 8 that all parameters are positively correlated with WQI, and Potassium shows the highest correlation with WQI.

Data pre-processing

It is essential to ensure data quality that should be checked before applying AI-learning algorithms. Normalization technique is applied to change the values of numeric columns and to utilize a common scale without losing information and distorting differences in the ranges of values.

- Min-max scaler: It is applied to scale the value within [0,1] and to compress all the inliers in the narrow range [0, 0.005]. In the presence of outliers, Standard Scaler does not guarantee balanced feature scales due to the influence of the outliers while computing the empirical mean and standard deviation. This leads to shrinkage in the range of the feature values.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{8}$$

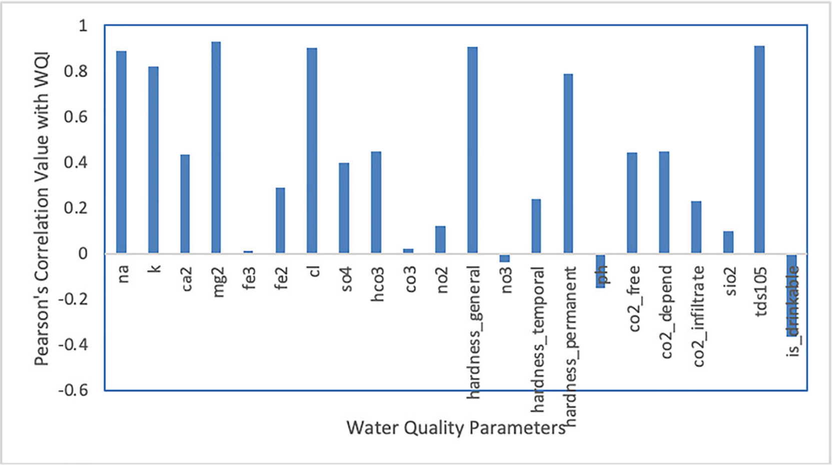


Fig. 7 Vietnam co-relation analysis of WQI with all other water parameters

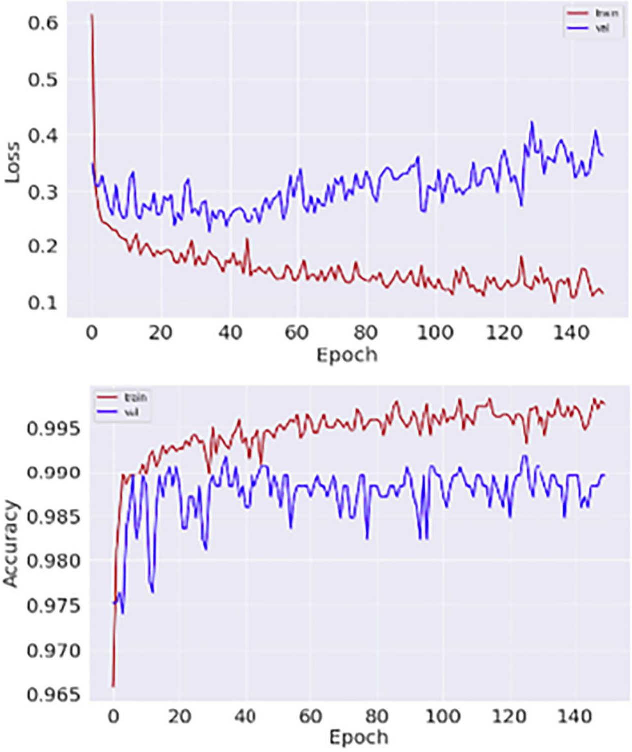


Fig. 8 Loss vs Accuracy for the Vietnam dataset

To deal with this matter, feature-wise normalization using as Min-Max Scaling as given in equation(7) is used before model-fitting.

Using machine learning models and performance evaluation

This section briefly describes the background of some state-of-the-art machine learning models which are applied for predicting water-quality classes based on drinkability. The models are then evaluated by using standard performance metrics.

The Models Hyperparameters and model training. Once data are preprocessed, we define the classifier for a given algorithm such as XGBoost, and then we fit and transform the classifier with the input trained data. We build the model using the classifier. We train the model using the train component of the data. Once the model is successfully trained, we evaluate the model's performance using the test component of the data. Hyperparameters of the applied models are used to build the model specific to the algorithm. Most of the cases, these Hyperparameters are used by default only.

The models are then evaluated by using standard performance metrics described in the hyperparameter table of our model in Table 11.

Models background

Logistic regression It is a classification algorithm based on supervised learning. It uses the sigmoid function or logit function. It fits the data into a logit function, which fits the line into a curve between 0 and 1. The line acts like an asymptotic line for the sigmoid curve. This model is mostly used for binary classification problems. But, to deal with multiple classes in our problem, a variant of this model is used, which is known as multinomial logistic regression. This model variant is followed in (Theyazn et al. 2020) to deal with multiple classes. The most important application of LR is to estimate the probability of the occurrence of an event, given information about predictors that may influence the outcome. (Hosmer and Lemeshow 1989), and (George and Meshack 2019), LRMs are distinguished from ordinary linear regression models as a class of generalized linear models by the range of their predicted values, the assumption of the variance of the predicted response, and the distribution of their prediction errors.

K-Nearest Neighbors K-Nearest Neighbors (K-NN) is a supervised classifier and non-parametric in nature. (Hmoud Al-Adhaileh and Waselallah Alsaade 2021) It uses the notion of data point proximity using distance computation to create a grouping of similar data points. Though it can be used both for classification and regression, it is mostly used as a classifier. In our case, Euclidean distance is used for determining the closeness of data points to different centroids. In case of a new point, it classifies it using the majority vote of K of its neighbours. The important and challenging step in the K-NN method is determining the optimal K value. In our case, the optimal K value is decided by plotting error Vs. K values as shown in Fig. 6. The decided K value is 2 and 3 as it gives the minimum error. In data science and machine learning, classification is a crucial issue. The KNN is one of the oldest and most accurate algorithms for pattern classification and regression models; it's easy to understand and implement for both classification and regression problems; it's ideal for non-linear data because it makes no assumptions about the underlying data; and it can naturally handle multi-class cases and perform well with sufficient representative data.

Decision Tree It is a hierarchical classification and a regression model in machine learning. By considering an instance from the samples, it makes a traversal of the tree and performs a comparison with important features with some pre-determined branching statements. A most important feature is selected as root, and subsequent levels are generated based on splitting other features. (Tiyasha et al. 2021) Decision trees are used to solve classification problems and classify things according to their learning characteristics. In addition, you can use them to solve

Table 11 Hyperparameters of few models

Hyperparameter of XGboost	
Model-xgb	xgb.XGBClassifier
Objective	binary: logistic, multi:softprob
Random-state	42
Colsample-bytree	0.8
Learning-rate	1
Max-depth	10
N-estimators	40
N-jobs	1
Min-samples-leaf	3
Min-samples-split	15
Scale-pos-weight	1
Hyperparameter of CNN Model	
Metrics is equal	sparse-categorical-crossentropy
The input shape	16
The activation	relu and softmax
Epochs	150
Model.add(Desnse	60
Adam optimizer	used to optimize network performance
Hyperparameter of SVM Model	
Linear	svm.SVC(kernel=linear)
Decision-function-shape	ovo
Poly	svm.SVC kernel=poly
Degree	3
Sig	svm.SVCkernel=sigmoid
Hyperparameter of SVM Adaboost	
AdaBoostClassifier	
N-estimator	100
Base-estimator	None
Learning-rate	1
Random-state	1

problems involving regression or forecast continuous results based on unanticipated data.

Support Vector Machine (SVM) SVM is a support vector machine and is widely used as a classifier. It is most effective in high-dimensional space. Different variants of SVM are available, linear SVM, polynomial SVM, RBF SVM, and sigmoid SVM. All these variants are applied to our problem, and the performance of the best variant is presented. (Arabgol et al. 2015) In machine learning, the significance of support vector machines (SVM) has been demonstrated by the fact that SVM can handle classification and regression on linear and non-linear data. (Nur Najwa Mohd et al. 2022) Adopted an SVM to predict the concentration and distribution of nitrate in groundwater.

AdaBoost AdaBoost is otherwise known as the adaptive boosting method. It is an ensemble technique used in machine learning. The adaptiveness lies in the reassignment

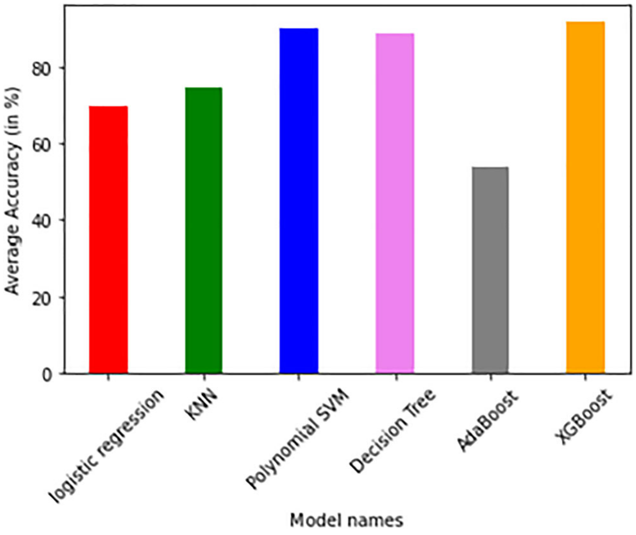


Fig. 9 Average accuracy of prediction of water quality classes for all models

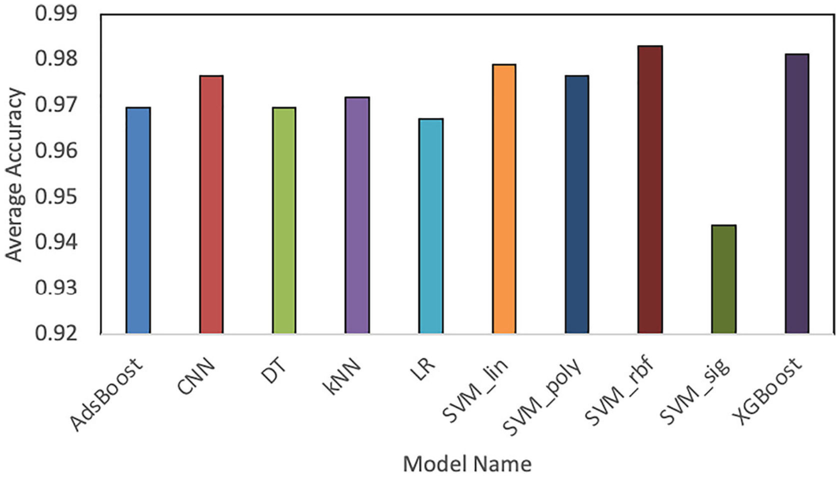
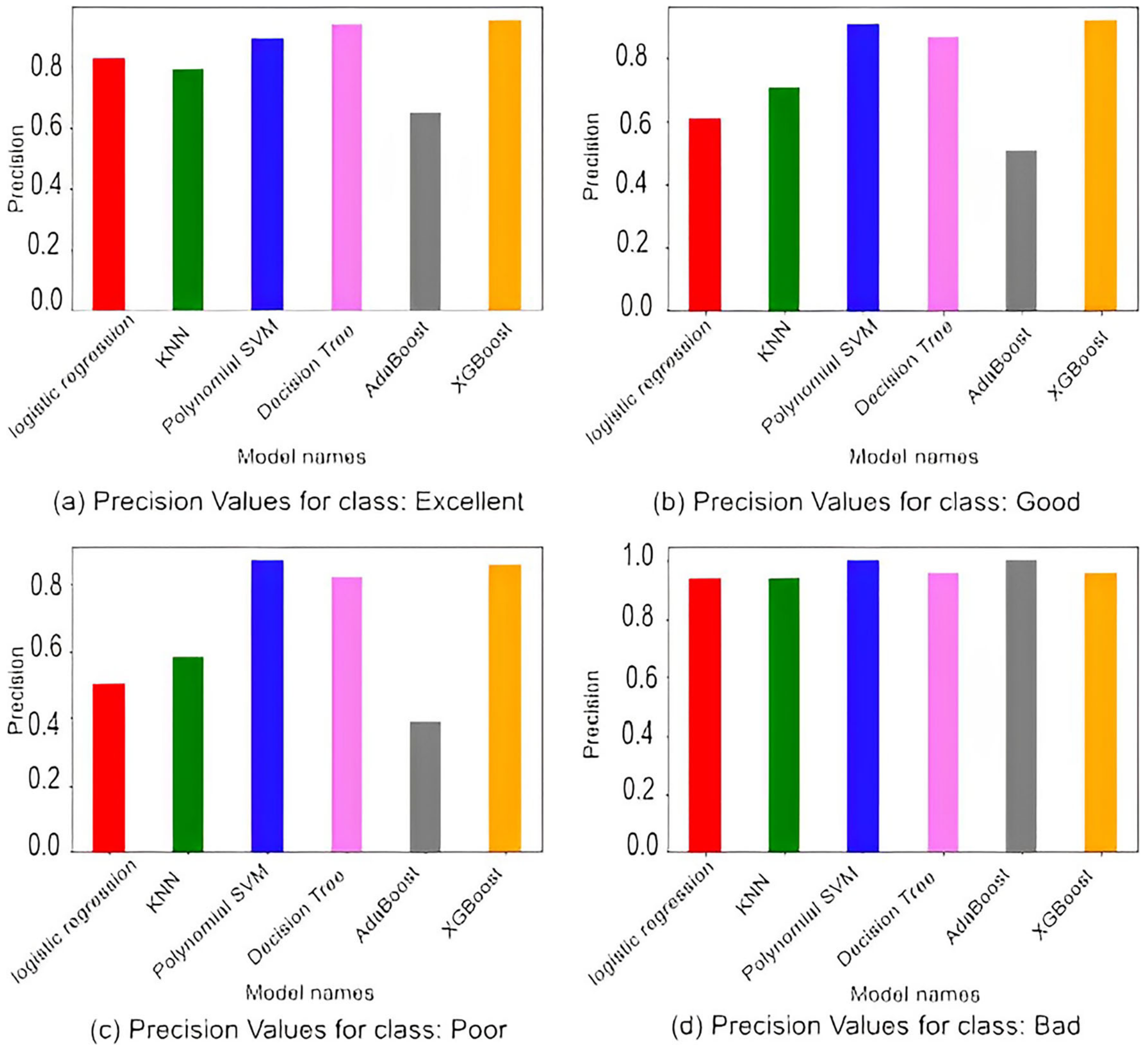


Fig. 10 Vietnam Average accuracy of prediction of water quality classes for all models

of weights to each instance in case of incorrectly classified instances. To reduce the bias, boosting is used.(Tu et al. 2017) AdaBoost has the benefits of being quick, easy to operate, and simple to program. Except for the number of iterations, no parameter adjustment is necessary. Without prior knowledge of WeakLearn, it can be combined flexibly



(a) Precision Values for class: Excellent

(b) Precision Values for class: Good

(c) Precision Values for class: Poor

(d) Precision Values for class: Bad

Fig. 11 Precision comparison of different models for Water Quality Classes

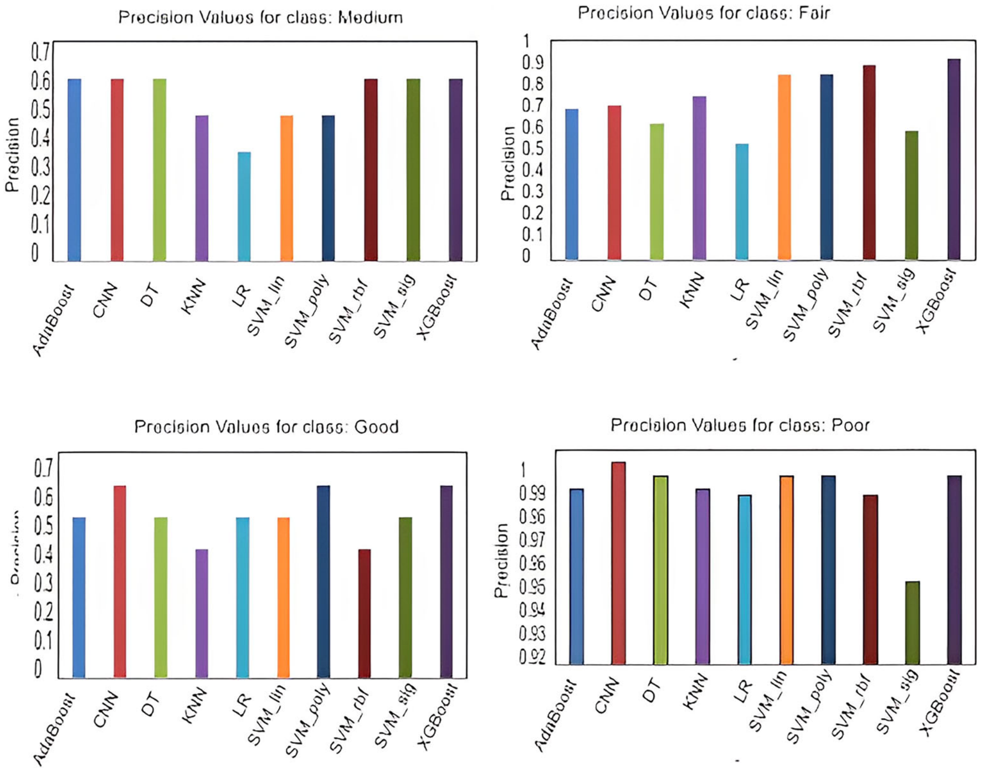
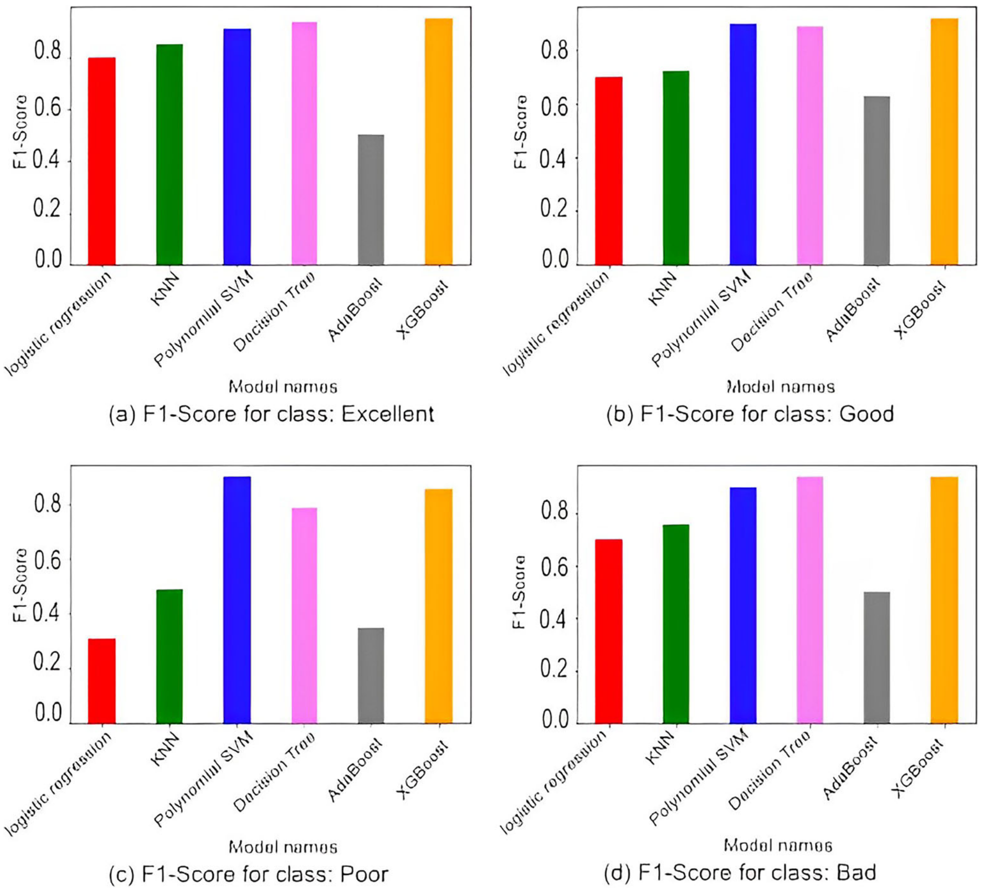


Fig. 12 Vietnam Precision comparison of different models for Water Quality Classes



(a) F1-Score for class: Excellent

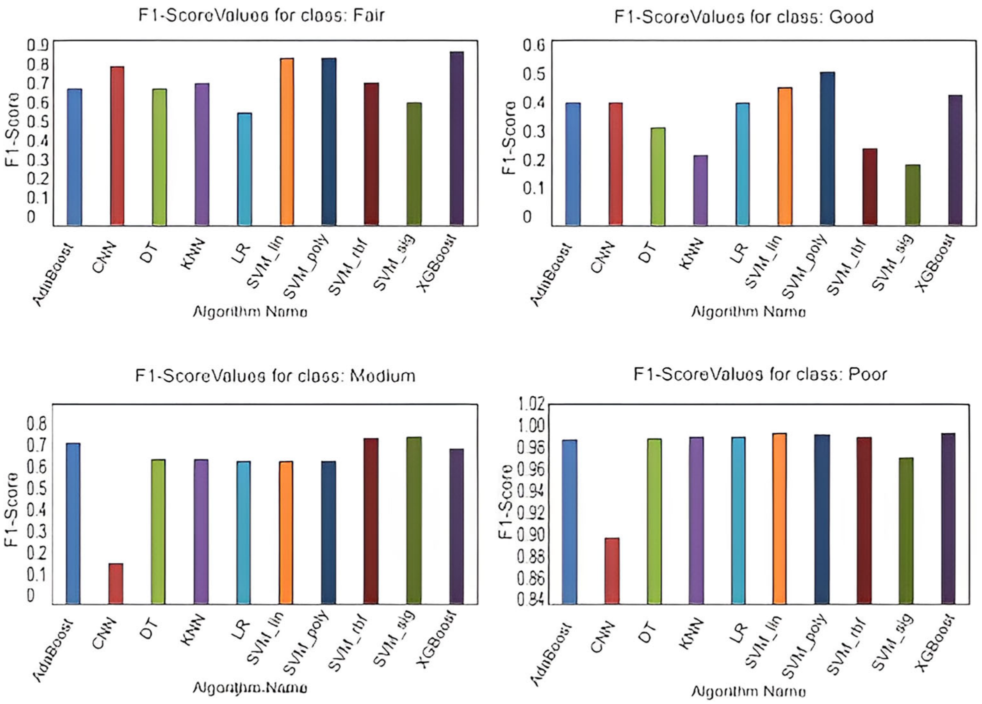
(b) F1-Score for class: Good

(c) F1-Score for class: Poor

(d) F1-Score for class: Bad

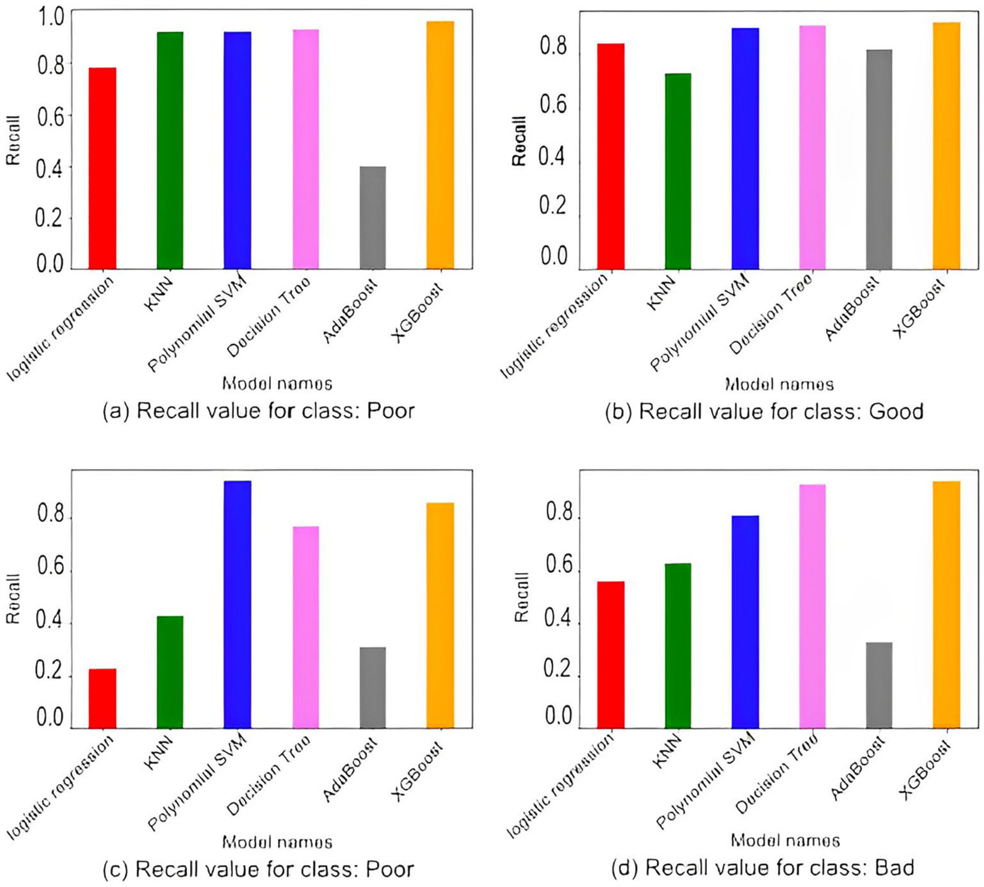
Fig. 13 F1-Score comparison of different models for Water Quality Classes

Fig. 14 F1-Score Vietnam comparison of different models for Water Quality Classes



with any method to seek a weak hypothesis. Given sufficient data and a WeakLearn with only reliable moderate accuracy, it can provide theoretical learning guarantees. (Alaa 2018) Adaboost is less susceptible to overfitting because the input

Fig. 15 Recall comparison of different models for Water Quality Classes Odisha



(a) Recall value for class: Poor

(b) Recall value for class: Good

(c) Recall value for class: Poor

(d) Recall value for class: Bad

Table 12 Comparison of the average value of performance metrics of all models Odisha

Model	Avg. Accuracy	Avg. Precision	Avg. F1-score	Avg. Recall
Logistic Regression	0.7051	0.7200	0.6275	0.6025
K-NN	0.7509	0.7550	0.7050	0.6775
Polynomial SVM	0.9012	0.9175	0.9025	0.8925
Decision Tree	0.8989	0.8975	0.8900	0.8850
AdaBoost	0.5445	0.6375	0.4950	0.4650
XGBoost	0.9267	0.9225	0.9175	0.9200

parameters are not jointly optimized. Using Adaboost, the accuracy of weak classifiers can be enhanced.

XGBoost XGBoost stands for Extreme Gradient Boosting. It is an ensemble model based on the tree concept. It is an enhancement of the gradient boosting framework by applying approximation algorithms. It provides a parallel tree-based framework for boosting. It applies to classification, regression and ranking problems. (Ramraj et al. 2016) XGBoost offers a few technical advantages over other gradient boosting approaches, including a more direct route to the minimum error, converging more quickly with fewer steps, and simplified calculations to improve speed and lower compute costs. In XGBoost, individual trees are created using multiple cores and data is organized to minimize the lookup times. This decreased the training time of models, which in turn increased the performance.

Convulutional neural network CNN CNNs have convolutional, pooling, and fully-connected layers. Local connections, shared weights, pooling, and layers are fundamental CNN applications. The fully-connected layers get their outputs. Backpropagation trains filter weights. (O'Shea and Nas 2015) Adam optimizer is used to optimize network performance. We utilized ReLU for the first and second layers and Softmax for the third. ReLU is a nonlinear function that outputs positive input directly and returns zero otherwise. CNN is the most widely used DL model in computer vision, image processing, speech recognition, natural language processing, and anomaly detection for drinking water using the BiLSTM ensemble technique(Chen et al. 2018). Significant importance of the CNNs model, the model does not require human supervision for the task of identifying essential features. They are very accurate at image recognition and classification. Weight sharing is another significant advantage of CNNs.

Performance metrics

The standard metrics used for evaluating the models are briefly presented below.

- *Accuracy*: It is measured in terms of the number of correct predictions done by the model over a total number of observed values. The corresponding equation is given in (9), where TP stands for true positive, TN stands for true negative, FP stands for false positive, and FN stands for false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

- Precision defines the ratio of correctly classified instances of a given class to the total classified instances of that particular class. It is calculated as given in equation (10).

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

- Recall is estimated as given in equation (11),

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Further, precision and recall alone can not reflect all aspects of the accuracy. Thus, the harmonic mean, i.e., F1-

Table 13 Comparison of the average value of performance metrics of all models Vietnam

Model	Avg. Accuracy	Avg. Precision	Avg. F1-score	Avg. Recall
Logistic Regression	0.9672	0.5333	0.5517	0.5714
K-NN	0.9719	0.9854	0.9902	0.9950
Polynomial SVM	0.9766	0.9902	0.9926	0.9950
Decision Tree	0.9696	0.9901	0.9889	0.9877
AdaBoost	0.9696	0.9853	0.9877	0.9901
CNN	0.9766	0.9950	0.9913	0.9877
XGBoost	0.9813	0.9902	0.9938	0.9975

score, is also computed as shown in equation (12). Its value lies between 0 and 1. The higher value of the F1-score reflects better accuracy.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

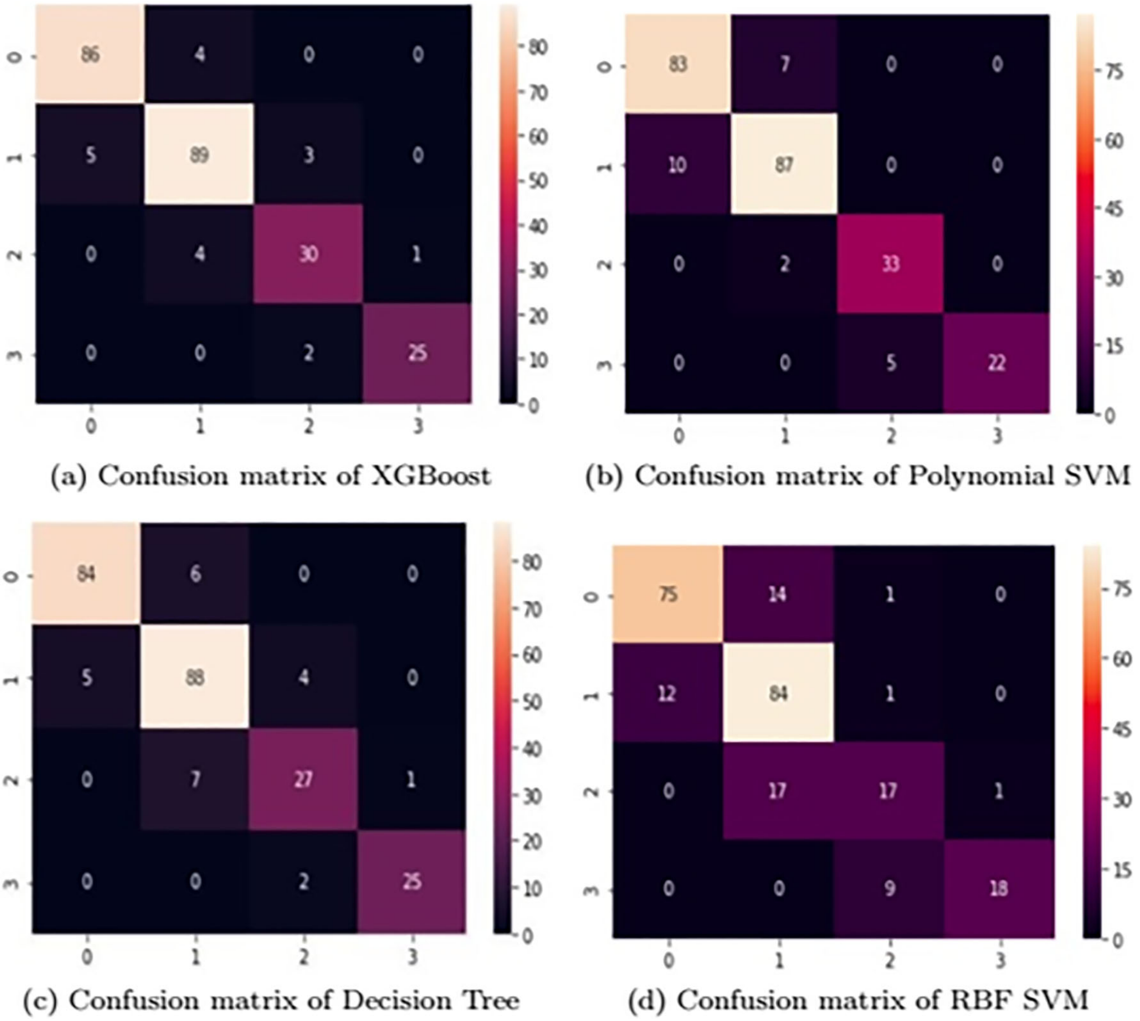
Finally, a confusion matrix is also found, which is an N × N matrix to evaluate the performance of a classification model, where N is the number of target classes. In our case, N is 4. The matrix compares the actual target value with the predicted value of the model. Hence, it gives a comprehensive view of the model's performance.

Performance of CNN

The plot for the loss and accuracy as a function of epochs for the Vietnam training and test sets to see how the network has performed is depicted below. The average accuracy achieved across classes is 99%.

Results summary & discussion

The performance of the above-applied models is summarized w.r. to average accuracy in Figs. 9 and 10 shows the average accuracy of classifying the samples according to all four classes. It is observed that XGBoost performance is optimal with an average accuracy of 92.67 % and 98%, followed by polynomial SVM with 90.3% and 97%. The average accuracy of the decision tree is 89.89% and 96%. Logistic regression and k-NN have an average accuracy of 70.51%, 75.09% and 907%, respectively. Poor performance is observed on Adaboost, with an average accuracy of 54.45% on the Odisha dataset. Using CNN on the Vietnamese dataset, the average accuracy across classes is 97%, while XGBoost performs optimally with an average accuracy of 98.13%, followed by polynomial SVM with 97.66%. The decision tree's average accuracy is 96.89%. The average accuracy of logistic regression and k-NN is 96.6% and 97.19%, respectively. In contrast to the poor performance



(a) Confusion matrix of XGBoost

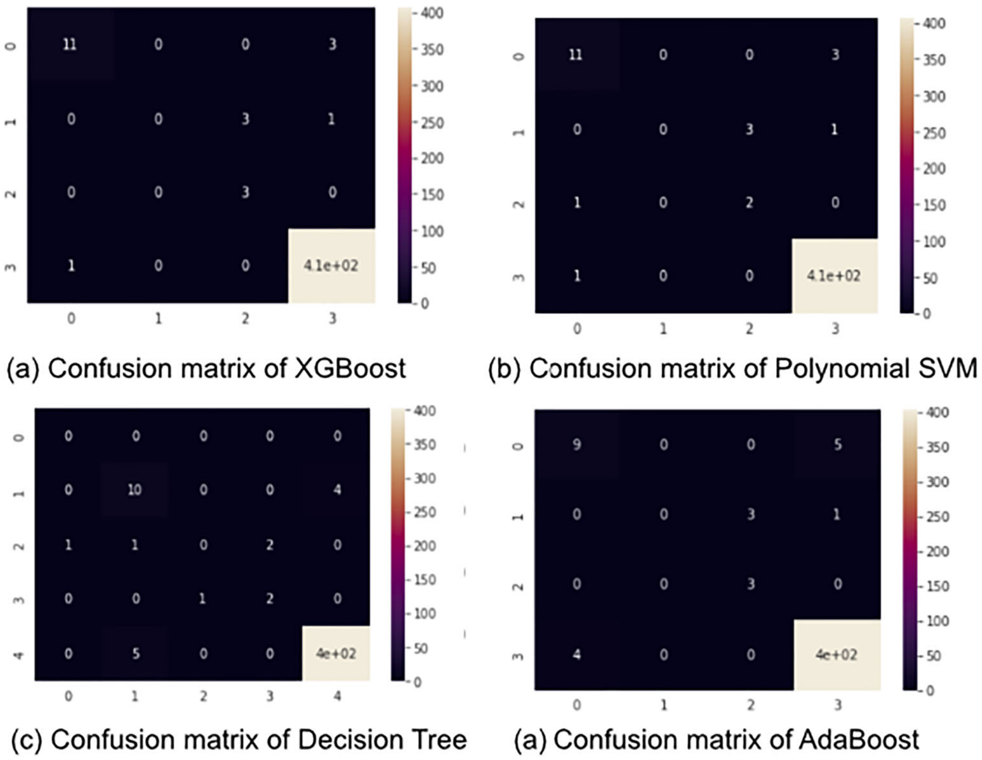
(b) Confusion matrix of Polynomial SVM

(c) Confusion matrix of Decision Tree

(d) Confusion matrix of RBF SVM

Fig. 16 Confusion matrices of some selected models for water quality classes of Odisha

Fig. 17 Confusion matrices of some selected models for water quality classes of Vietnam dataset



(a) Confusion matrix of XGBoost (b) Confusion matrix of Polynomial SVM

(c) Confusion matrix of Decision Tree (a) Confusion matrix of AdaBoost

Table 14 Performance metrics of Logistic regression model

Water Class	Precision	F1-score	Recall
Performance measures of Odisha groundwater data			
Excellent	0.83	0.80	0.78
Good	0.61	0.70	0.84
Poor	0.50	0.31	0.23
Bad	0.94	0.70	0.56
Performance measures of Vietnam groundwater data			
Good	0.71	0.80	0.61
Medium	0.80	0.78	0.83
Poor	0.98	0.99	0.99
Fair	0.53	0.55	0.57

Table 15 Performance metrics of K-NN model

Water Class	Precision	F1-score	Recall
Performance measures of Odisha groundwater data			
Excellent	0.79	0.85	0.92
Good	0.71	0.72	0.73
Poor	0.58	0.49	0.43
Bad	0.94	0.76	0.63
Performance measures of Vietnam groundwater data			
Good	0.89	0.85	0.90
Medium	0.41	0.67	0.50
Poor	0.99	1	0.99
Fair	0.75	0.64	0.69

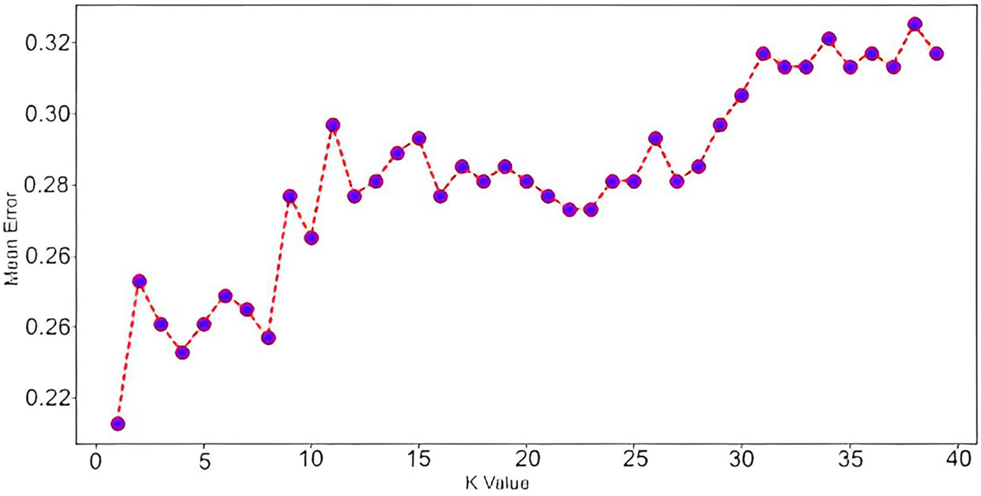


Fig. 18 Mean error vs. K-value

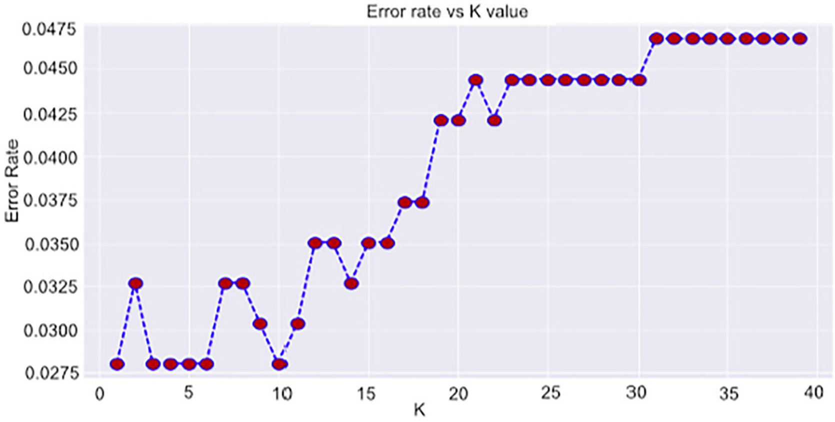


Fig. 19 Mean error vs. K-value

Table 16 Performance metrics of Polynomial SVM

Water Class	Precision	F1-score	Recall
Performance measures of Odisha groundwater data			
Excellent	0.89	0.91	0.92
Good	0.91	0.90	0.90
Poor	0.87	0.90	0.94
Bad	1.00	0.90	0.81
Performance measures of Vietnam groundwater data			
Good	0.78	0.90	0.92
Medium	0.4	0.5	0.66
Poor	0.99	0.99	0.99
Fair	0.84	0.81	0.78

Table 17 Performance metrics of Decision Tree

Water Class	Precision	F1-score	Recall
Performance measures of Odisha groundwater data			
Excellent	0.94	0.94	0.93
Good	0.87	0.89	0.91
Poor	0.82	0.79	0.77
Bad	0.96	0.94	0.93
Performance measures of Vietnam groundwater data			
Good	0.5	0.33	0.25
Medium	0.5	0.57	0.66
Poor	0.99	0.98	0.98
Fair	0.62	0.66	0.71

observed with Adaboost on the Odisha dataset, with an average accuracy of 54.45%, the model performed better on the Vietnamese dataset, with an average accuracy of 96.96%.

Figures 11 and 12 show the water quality class-wise precision comparison of all applied models. It is observed that for the considered dataset, XGBoost is able to classify with the highest precision value for classes, Excellent and Good, with an average precision of 0.9225 and 0.9813. Similar performance is observed from polynomial SVM and decision tree for all classes with an average precision of 0.9175 and 0.8975, respectively. The average precision for Adaboost is 0.6375, which is the lowest among all models comparison in Odisha. The other performance metrics, including precision, F1-score and recall, are compared class-wise.

The comparison of performance among six models by Precision is shown in Fig. 11 and 12 below.

The comparison of performance among six models by F1-Score is given in Figs. 13 and 14.

The Recall values obtained by applying these models are compared in Fig. 15.

In Fig. 13, a similar process is followed as in Fig. 11, but it represents the class-wise F1-score comparison among the six used models. By averaging all four classes' F1-score values, it is observed that XGBoost has an average F1-score of 0.9175% and 0.9938%, followed by 0.9025 and 0.9926 for polynomial SVM while 0.8900 and 0.9889 for decision tree. The average F1-score for AdaBoost is the lowest among all applied models, with the value 0.4950 on the Odisha dataset while 0.9877 is obtained for the Vietnam data. Figure 15 shows a recall value comparison between the applied models for all four classes. Similar observations are made on average recall value. XGBoost performs optimally with an average recall value of 0.9200 and 0.9975. In contrast, Adaboost performs poorly on the sample dataset, with an average recall value of 0.4650 on Odisha data and 0.9901 on Vietnam data. The average value of all performance metrics is shown in Tables 12 and 13. It is observed that XGBoost and polynomial SVM have shown optimal performance.

Table 18 Performance metrics of AdaBoost

Water Class	Precision	F1-score	Recall
Performance measures of Odisha groundwater data			
Excellent	0.65	0.50	0.40
Good	0.51	0.63	0.82
Poor	0.39	0.35	0.31
Bad	1.00	0.50	0.33
Performance measures of Vietnam groundwater data			
Good	0.5	0.33	0.25
Medium	0.5	0.57	0.66
Poor	0.99	0.98	0.98
Fair	0.62	0.66	0.71

Further, the confusion matrices of some selected models such as XgBoost, RBF SVM, polynomial SVM and decision tree are presented in Figs. 16 and 17 to make a holistic observation about the model performance. The results show that the XgBOOST and Polynomial SVM model accurately classified the water quality classes with an accuracy of 92% and 90%, respectively, on Odisha data. In contrast, Vietnam data has 97% and 98%, respectively.

Performance of logistic regression

Logistic regression worked with an accuracy of 70 % on Odisha data and 96% on Vietnam data with other performance metrics as shown in Table 14. The reason for less accuracy may be caused by less correlation between the parameters, which does not help well the Logistic regression to classify.

Performance of K-Nearest Neighbor

The supervised KNN model is also applied for the prediction of the water classes as shown in Table 15. To find the best value of K, we plot the graph between Error and K Value. Based on the value of K in which the error is minimum, we select the best K value. We find 2,3 as the best values of K which resulted in an accuracy of 75 %. Other performance metrics are shown in Table 15 and on the Vietnamese dataset to find the best value of K, and we plot the graph between Error Rate and K Value. The error rate is found to be minimum at K=10, thus resulting in an accuracy of 97%. Error Rate with K is shown in Figs. 18 and 19.

The dependency of errors on the value of K is also given in Figs. 18 and 19.

Table 19 Performance metrics of XGBoost

Water Class	Precision	F1-score	Recall
Performance measures of Odisha groundwater data			
Excellent	0.95	0.95	0.96
Good	0.92	0.92	0.92
Poor	0.86	0.86	0.86
Bad	0.96	0.94	0.94
Performance measures of Vietnam groundwater data			
Good	0.96	0.95	0.96
Medium	0.50	0.67	1.00
Poor	0.99	0.99	0.99
Fair	0.69	0.67	0.64

Performance of support vector machine and its variants

The following variants of SVM are applied and the best-performing SVM variant's result is summarized. Variants of SVM

Linear SVM

This is applicable to linearly separable problems. We performed linear SVM which resulted in an accuracy of 75.5%. on Odisha and 97% on Vietnam dataset

Polynomial SVM

Polynomial Kernel: The polynomial kernel features are added with higher-order polynomials to analyze the data. This mechanism is implemented by the SVC class and the accuracy obtained in Odisha is 90.3 % and 97%. Table 16 show the metrics obtained by polynomial SVM model for Odisha and Vietnam.

RBF SVM

The RBF kernel function is used in this case with two hyper-parameters: (i) Gamma, and (ii) C (regularization parameter). A lower C value is set at the cost of training accuracy of on Odisha 77.9% and 97% on the Vietnam dataset.

Sigmoid SVM

The sigmoid kernel function is used with an accuracy of 27.3%.

Among the four variants of SVM, the polynomial SVM shows the maximum accuracy on Odisha (90 %) and 94% on the Vietnam dataset. Table 13 gives results of other performance measurements.

Performance of Decision Tree

The decision tree model gives the accuracy as 89% on Odisha and 96% on the Vietnam dataset, and other performance metrics are shown in Table 17.

Performance of AdaBoost

The Adaptive Boosting algorithm (Adaboost) is an ensemble method of learning. The accuracy obtained on Odisha is 54 % and 96% on the Vietnam dataset with other performance metrics as shown in Table 18.

Performance of XGBoost

XGBoost is a hierarchy-based ensemble machine learning algorithm. The accuracy obtained in Odisha is 92 % and 98% in Vietnam. Table 19 shows other performance metrics obtained by applying this model.

Conclusion

Groundwater quality monitoring is an important prerequisite for water management. In this article, a case study on the state of Odisha, India and Vietnam's northern delta is conducted to predict the water quality for drinking purposes. As a first step, exploratory data analysis is used to remove non-contributing parameters, $CO3^-$, to analyze the water quality dataset, to detect outliers and to find out the correlation between different water quality parameters. Further, a set of representative supervised AI-learning algorithms are used to compute WQI. The water metrics, including pH, EC, Total Phenol, Harshness - permanent, TDS, Turbidity, Chloride, Magnesium, Sodium, Chloride, Alkalinity etc, were used in this study. For the classification purpose, we have used different models such as Logistic regression, KNN, CNN, AdaBoost, XGBoost, SVM and its variants, and decision tree where XGBoost and Polynomial SVM worked well with an accuracy of 92% and 98 %, respectively. The average accuracy across classes for CNN on the Vietnam data is 99%.

The present work uses 13 water quality parameters from Odisha and 12 parameters from Vietnam which can be further minimized, and a lesser number of parameters can be used to predict the drinkability class without compromising the accuracy. The scope of work can also be extended to validate the performance of Machine Learning models by applying them to similar datasets collected from other countries. Further, the WQI ranges can be fuzzified, and a fuzzy inference system can be developed for predicting the water classes.

Acknowledgements We thank the three referees for their suggestions which improved the manuscript.

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Michael Omar, Tran Thi Ngan, Bui Thi Thu, Raghvendra Kumar and Niranjan Panigrahi. The first draft of the manuscript was written by Michael Omar, S. Gopal Krishna Patro, Nguyen Long Giang and Nguyen Truong Thang. All authors read and approved the final manuscript.

Funding This research has been funded by the Research Project: CSCL02.03/23-24, Institute of Information Technology, Vietnam Academy of Science and Technology. We are grateful for the support from the staff of the Institute of Information Technology, Vietnam Academy of Science and Technology.

Data Availability You can contact Michael Omar (omar2@fe.edu.vn) for data and materials.

Declarations

Conflict of Interests The authors declare no conflict of interest

References

Ahmed U, Mumtaz R, Anwar H, Mumtaz S, Qamar AM (2020) Water quality monitoring: from conventional to emerging technologies. Water Supply 20(1):28–45. https://doi.org/10.2166/ws.2019.144

Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J (2019) Efficient water quality prediction using supervised machine learning. Water 11:2210. https://doi.org/10.3390/w11112210

Alaa T (2018) AdaBoost classifier: an overview. https://doi.org/10.13140/RG.2.2.19929.01122

Arabgol R, Sartaj M, Asghari K (2015) Predicting nitrate concentration and spatial distribution in groundwater resources using support vector machines (SVMs) model. Environ Model Assess 21:71–82. https://doi.org/10.1007/s10666-015-9468-0

Barik R, Pattanayak SK (2019) Assessment of groundwater quality for irrigation of green spaces in the Rourkela city of Odisha, India. Groundwater Sustain Dev 8:428–438. https://doi.org/10.1016/j.gsd.2019.01.005

Chen X, Feng F, Wu J, Liu W (2018) Anomaly Detection for Drinking Water Quality via Deep biLSTM Ensemble. In: Proceedings of the genetic and evolutionary computation conference companion, edited by Hernan Aguirre. ACM, pp 3–4, New York

Dogo EM, Nwulu NI, Twala B, Aigbavboa C (2019) A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J 16(3):235–248. https://doi.org/10.1080/1573062X.2019.1637002

George K, Meshack A (2019) Groundwater quality prediction using logistic regression model for Garissa County

Haghiabi AH, Nasrolahi AH, Parsaie A (2018) Water quality prediction using machine learning methods. Water Qual Res J 53(1):3–13. https://doi.org/10.2166/wqrj.2018.025

Ground Water Year Book (2018). Available on http://cgwb.gov.in/gw-yearbook-state.html

Hanoon MS, Ahmed AN, Fai CM et al (2021) Application of artificial intelligence models for modeling water quality in groundwater: comprehensive review, evaluation and future trends. Water Air Soil Pollut 232:411. https://doi.org/10.1007/s11270-021-05311-z

Harichandan A, Patra HS, Dash AK et al (2021) Suitability of groundwater quality for its drinking and agricultural use near Koira region of Odisha, India. Sustain Water Resour Manag 7:51. https://doi.org/10.1007/s40899-021-00505-z

Hmoud Al-Adhaileh M, Waselallah Alsaade F (2021) Modelling and prediction of water quality by using artificial intelligence. Sustainability 13(8):4259. https://doi.org/10.3390/su13084259

Hosmer DW, Lemeshow S (1989) Applied logistic regression. Wiley, New York

Khan SI, Islam N, Uddin J, Islam S, Nasir MK (2021) Water quality prediction and classification based on principal component regression and gradient boosting classifier approach, J King Saud Univ - Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2021.06.003

Khan Y, See CS (2016) Predicting and analyzing water quality using Machine Learning: a comprehensive model. In: 2016 IEEE Long island systems, applications and technology conference (LISAT), pp 1-6. https://doi.org/10.1109/LISAT.2016.7494106

Kouadri S, Elbeltagi A, Islam ARMT et al (2021) Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl Water Sci 11:190. https://doi.org/10.1007/s13201-021-01528-9

Madhav S, Kumar A, Kushawaha J et al (2020) Geochemical assessment of groundwater quality in Keonjhar City, Odisha, India. Sustain Water Resour Manag 6:46. https://doi.org/10.1007/s40899-020-00395-7

Nur Najwa Mohd R, Hayder G, Yusof KA (2022) Water quality predictive analytics using an artificial neural network with a graphical user interface. Water 14(8):1221. https://doi.org/10.3390/w14081221

O'Shea K, Nas R (2015) An introduction to convolutional neural networks, arXiv:1511.08458v2

Ramraj S, Uzir N, Sunil R, Banerjee S (2016) Experimenting XGBoost algorithm for prediction and classification of different datasets. Int J Control Theory Applic 9(40):651–662

Ranjithkumar M, Robert L (2021) Machine Learning Techniques and Cloud Computing to Estimate River Water Quality-Survey. In: Ranganathan G, Chen J, Rocha Á (eds) Inventive communication and computational technologies. lecture notes in networks and systems. Springer, 145, Singapore. https://doi.org/10.1007/978-981-15-7345-3_32

Sahu S, Gogoi U, Nayak NC (2021) Groundwater solute chemistry, hydrogeochemical processes and fluoride contamination in phreatic aquifer of Odisha, India Geosci. Geosci Front 12(3):101093. https://doi.org/10.1016/j.gsf.2020.10.001

Sokolova E, Ivarsson O, Lillieström A, Speicher NK, Rydberg H, Bondelind M (2022) Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data. Sci Total Environ 802:149798. https://doi.org/10.1016/j.scitotenv.2021.149798

Theyazn H, Aldhyani H, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artificial intelligence algorithms. Appl Bionics Biomech 2020:12. https://doi.org/10.1155/2020/6659314

Tiyasha T, Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. J Hydrol 585:124670. https://doi.org/10.1016/j.jhydrol.2020.124670

Tiyasha T, Tung TM, Yaseen ZM (2021) Deep learning for prediction of water quality index classification: tropical catchment environmental assessment. Nat Resour Res 30:4235–4254. https://doi.org/10.1007/s11053-021-09922-5

Tiyasha T, Tung TM, Bhagat SK, Tan ML, Jawad AH, Wan Mohtar WHM, Yaseen ZM (2021) Functionalization of remote sensing and on-site data for simulating surface water dissolved oxygen: development of hybrid tree-based artificial intelligence models. Mar Pollut Bull 170:112639. https://doi.org/10.1016/j.marpolbul.2021.112639

Tu C, Liu H, Xu B (2017) AdaBoost typical Algorithm and its application research. MATEC Web Conferences 139:00222. https://doi.org/10.1051/matecconf/201713900222

Ubah JI, Orakwe LC, Ogbu KN, Awu JI, Ahaneku IE, Chukwuma EC (2021) Forecasting water quality parameters using artificial neural network for irrigation purposes. Sci Rep 11(1):24438. https://doi.org/10.1038/s41598-021-04062-5

Venkata Vara Prasad D, Senthil Kumar P, Venkataramana LY, Prasannamedha G, Harshana S, Jahnavi Srividya S, Harrinei K, Indraganti S (2021) Automating water quality analysis using ML and auto ML techniques. Environ Res 202:111720. https://doi.org/10.1016/j.envres.2021.111720

Venkata Vara Prasad D, Venkataramana LY, Senthil Kumar P, Prasannamedha G, Harshana S, Jahnavi Srividya S, Harrinei K, Indraganti S (2022) Analysis and prediction of water quality using deep learning and auto deep learning techniques. Sci Total Environ 821:153311. https://doi.org/10.1016/j.scitotenv.2022.153311

WHO Guidelines for drinking water quality (2004). https://cpcb.nic.in/who-guidelines-for-drinking-water-quality/. Accessed 31 May 2022

Wang R, Kim JH, Li MH (2021) Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Sci Total Environ 761:144057. https://doi.org/10.1016/j.scitotenv.2020.144057

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. Sci Rep 7:12858. https://doi.org/10.1038/s41598-017-12853-y

Zhu M, Wang J, Yang X, Zhang Y, Zhang L, Ren H, Wu B, Ye L (2022) A review of the application of machine learning in water quality evaluation. Eco-Environ Health 1(2):107–116. https://doi.org/10.1016/j.eehl.2022.06.001

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Niranjan Panigrahi[1] · S. Gopal Krishna Patro[2] · Raghvendra Kumar[3] · Michael Omar[4] · Tran Thi Ngan[5] · Nguyen Long Giang[6] · Bui Thi Thu[6] · Nguyen Truong Thang[6]

Niranjan Panigrahi
niranjan.cse@pmec.ac.in

S. Gopal Krishna Patro
sgkpatro2008@gmail.com

Raghvendra Kumar
raghvendraagrawal7@gmail.com

Michael Omar
omar2@fe.edu.vn

Tran Thi Ngan
ngantt@tlu.edu.vn

Bui Thi Thu
minhthuph21@gmail.com

Nguyen Truong Thang
ntthang@ioit.ac.vn

[1] Department of CSE, PMEC, Berhampur, Odisha, 761003, India

[2] Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 10587, Andhra Pradesh, India

[3] Department of CSE, GIET University, Gunpur, Odisha, 610101, India

[4] Department of Information Technology, FPT Greenwich, Hanoi, 100000, Vietnam

[5] Faculty of Computer Science and Engineering, Thuyloi University, Hanoi, 100000, Vietnam

[6] Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, 100000, Vietnam