Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 22 May 2026
Revised 25 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 35–58

KEYWORDS:
Groundwater Quality Assessment, Drinkability Prediction, Machine Learning, Water Quality Index, Hydrochemical Parameters, Multi-Class Classification, XGBoost, Support Vector Machine.

Machine Learning-Based Groundwater Quality Assessment and Drinkability Prediction

[1]Onkar Nath Thakur, [2]Santosh Kumar, [3]Nitesh Singh Bhati

[1]School of Computing Science & Engineering, Galgotias University, Greater Noida
er.onkarthakur@gmail.com
[2]School of Computing Science & Engineering, Galgotias University, Greater Noida
Sant7783@hotmail.com
[3]School of Computing Science & Engineering, Gautam Buddha University, Greater Noida
Niteshbhati07@gmail.com

ABSTRACT: Groundwater is a major source of drinking water for a large proportion of the global population; however, rapid urbanization, industrialization, and agricultural activities have led to a marked deterioration of groundwater quality. Thus, there is a need for efficient, reliable and rapid groundwater quality assessment techniques for identifying contamination hotspots. The typical measurement of groundwater quality involves physicochemical analysis and WQI calculation, which is a time-taking and labor-intensive method and not suitable for large-scale monitoring. This study applies the supervised machine learning method for the automated assessment of groundwater quality and prediction of its drinkability quality in terms of hydrochemical characteristics. The Central Ground Water Board (CGWB), India, contains a dataset of 1345 groundwater samples containing 13 hydrochemical parameters (pH; Electrical Conductivity (EC); Total Dissolved Solids (TDS); Total Hardness (TH); Alkalinity; Calcium; Magnesium; Sodium; Potassium; Chloride; Sulphate; Fluoride; Bicarbonate). The Water Quality Index (WQI) was computed based on standard guidelines for drinking water quality and groundwater samples were categorized into four drinkability classes viz. Excellent, Good, Poor and Bad thereby creating a multi-class classification problem. Before the development of the applied model, exploratory data analysis, correlation analysis, outlier detection, data cleansing, and feature normalization were done to improve data quality. Using accuracy, precision, recall, and F1-score six supervised machine learning algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), AdaBoost and Extreme Gradient Boosting(XGBoost) were comparatively evaluated. Consistent with previous findings, experimentation shows that ensemble and tree-based learning algorithms outperform standard classifiers in groundwater quality classification. Of the algorithms trained, XGBoost performed the best with a classification accuracy of 98%, followed by Polynomial SVM, whose accuracy was 97% and decision tree with 96%. The comparison provides benchmark performance for conventional machine learning models, which show potential for rapid, low-cost, data-driven monitoring of groundwater quality and drinking water management.

1. INTRODUCTION

Groundwater is one of the most important freshwater resources as nearly half of drinking water comes from it while some domestic, agricultural and industrial activities are supported. Groundwater is the major source of drinking water in many developing countries including India because of its easy availability and cheaper treatment cost. But rapid urbanization, industrialization, excessive agricultural practices and climate variability have adversely affected the groundwater quality through the build-up of dissolved salts, heavy metals, fluoride, nitrates and other hydrochemical pollutants. The declining quality of groundwater poses serious risks to public health, food security, and sustainable water resources management which has made continuous monitoring and accurate quality assessment a global research issue. The qualities exhibited in groundwater are results of geological formations, hydrogeochemical processes, man-made activities, and climatic conditions. As a result, it is not easily to assess quality due to the spatial and temporal variability of groundwater characteristics. In traditional evaluation of groundwater quality, the physicochemical analyses of hydrochemical parameters are carried out in a laboratory. Water Quality Index (WQI) is then extensively calculated. The WQI integrates multiple water quality parameters into a single numerical indicator representing the suitability of groundwater for drinking purposes. Water Quality Index (WQI) is an inherently good tool for supporting decision-making by the water resource managers. However, the computation of water quality index will require laboratory testing of numerous samples, expert knowledge, testing and auditing, a lot of time and money. WQI is thus not suitable for large scale implementation for real-time monitoring [11,12]. Advances in artificial intelligence (AI) and machine learning (ML) have enabled data-driven prediction of complex hydrogeochemical processes, and transforming environmental monitoring. Machine learning algorithms can model nonlinear relationships among hydrochemical variables without requiring a defined mathematical formulation unlike traditional statistical methods. The aforementioned capabilities have encouraged researchers utilizes supervised learning method in effectively predicting groundwater quality and classifying water quality, detection of contaminant, and further, Water Quality Index (WQI) [1–9]. A world-class review study has cited that machine learning models are far better than ordinary statistical and regression-based tools for multidimensional groundwater data and assure better predictive accuracy and computational efficiency [2]. Many recent studies show that machine learning is effective for groundwater quality assessment. Various artificial intelligence-based systems have been developed for groundwater quality assessment and assessment of the potability of water from hydrochemical parameters. The integration of machine learning, geospatial analysis, and Water Quality Index approaches has further improved groundwater quality assessment predictive ability [6]. Likewise, several supervised machine learning algorithms have been deployed to predict groundwater quality and irrigation water quality indices under varying hydrochemical conditions [9]. The application of machine learning models (ML) for the prediction of groundwater quality has gained momentum. The availability of abundant hydrochemical datasets has further supplemented this increase. Coastal groundwater quality prediction using Objective-weighted Water Quality Index methods integrated with machine learning techniques shows better classification reliability [7]. Similarly, water quality classification proposals with machine learning frameworks with numerous water quality categories has been conducted using supervised learning algorithm [17]. More recent researches, as mentioned in [8,10], combine machine learning models with geospatial technologies to that predict groundwater quality quantity and environmental monitoring. Taken together, these studies suggest that machine learning can significantly reduce computational burden and improve prediction abilities for groundwater quality assessment. Recent studies have also looked at machine learning mixed with GIS, hydrochemical analysis, and environmental modeling. Integrated machine learning frameworks that are based on Water Quality Index yield promising results in groundwater quality prediction in semi-arid river basins and drinking water suitability assessment [5]. The findings show that Machine Learning and GIS have shown the reliable prediction of groundwater quality for agricultural practice [14]. Similarly, hydrochemical investigations were carried out to study hydraulic analysis of fluoride enrichment [13]. Improved Water Quality Index analysis and spatial modeling was conducted for satisfactory nutrient index in groundwater [15]. The studies show that hydrochemical indicators combined with data-driven learning algorithms are becoming increasingly important for groundwater quality evaluation. Even with these developments, research challenges persist. To begin with, a large number of existing studies develop predictive models for continuous Water Quality Index values rather than useful drinkability classes. Moreover, multiple studies evaluate one or two machine learning algorithms only, which does not allow us to conclude the classifier that should best be used for groundwater quality assessment. Another challenge is that the hydrochemical characteristics of river water of specific regions differ from each other. So, they are not generalisable. Therefore, comparative evaluations may be done on regional datasets through multiple algorithms. In addition to this, while advanced deep learning and explainable artificial intelligence techniques have attracted significant attention recently, obtaining robust baseline performance using traditional supervised machine learning models should be done before developing more sophisticated predictive methods [2–6]. This study aims to assess the effectiveness of traditional supervised machine learning methods for predicting groundwater quality and its potability using hydrochemical features. Water Quality Index classification of groundwater quality by the use of machine learning techniques, these groundwater samples collected from Central Ground Water Board (CGWB), India. The WPQI standards classify ground water samples into classes like Excellent, Good, Poor, Bad which represents their drinkability. Before developing the model, exploratory data analysis, preprocessing of data, correlation analysis, outlier detection, and normalization of features are applied to enhance the quality of the data. In order to achieve this functionality, a comparative analysis is made possible among Logistic Regression, KNN, Decision Tree, SVM, AdaBoost, and XGBoost with the help of some metrics that provide classification efficiency – to name a few, accuracy, precision, recall, and F1 score. The key contributions of this study can be summarized as:

- A framework for groundwater quality assessment is prepared using hydrochemical parameters and Water Quality Index-based classification for drinkability.
- The prediction of drinkability of groundwater is framed by making a four-class supervised machine learning classification problem viz Excellent, Good, Poor, and Bad category of water.
- A comprehensive comparison of six popular supervised machine learning algorithms was done for selecting the appropriate model for groundwater quality prediction. Using a systematic comparative evaluation of six commonly used supervised machine learning algorithms is carried out to select the suitable model for groundwater quality prediction.
- The models' predictive performance is assessed based on several evaluation metrics so as to establish benchmark results for groundwater drinkability prediction.
- According to the results, a reliable basis is created for future works on advanced frameworks for eXplainable artificial intelligence and hybrid ensemble learning for groundwater quality.

2. RELATED WORKS

The groundwater quality assessment is gaining wide research interest with the rising demand for safe drinking water and negative impact of industrialization, urbanization and agricultural practices on groundwater resources. Traditionally, groundwater quality has been assessed through physicochemical analysis and Water Quality Index (WQI) models which aggregates several hydrochemical parameters into a single indicator reflecting suitable drinking and irrigation water quality. While WQI-based methods are extensively applied, the process requires extensive laboratory testing, data interpretation by experts, which cannot be adopted here for rapid large-scale assessment. Studies done recently have increasingly explored the use of machine learning techniques to automate groundwater quality assessment and improve prediction accuracy. Panigrahi et al. (1) in 2020 designed framework based on Artificial intelligence which can be used for groundwater quality analysis in India. These hydrochemical parameters were collected from CGWB (Central Ground Water Board), India. The present study had formulated the drinking water quality index of groundwater as a classification problem. In addition to this, the use of machine learning models can enhance accuracy. Despite this, the paper mainly focused on prediction performance and did not provide any comparisons between different supervised learning algorithms. A comprehensive review of machine learning applications in groundwater quality modelling is presented by Haggerty et al. [2]. The paper reviewed several supervised and unsupervised learning algorithms including ANN, SVM, DT, RF and ensemble learning. Tree-based ensemble models tend to outperform conventional regression models if the review is anything to go by. However, the authors stressed the importance of having standard benchmark studies on similar hydrochemical datasets. According to Huang et al. [3], advances in machine learning for groundwater quality evaluation and prediction were systematically reviewed. The various data pre-processing and feature engineering methods employed in predictive modelling are analysed along with the challenges of data scarcity, model generalization and interpretability. Torres-Martínez et al. [4], likewise analyzed state-of-the-art groundwater quality prediction methods, and reported that although models using machine learning achieve good prediction performance, little comparative evaluation of multiple algorithms exists, notably for regional groundwater datasets. Recent researchers worked on a multi-level Water Quality Index computation machine learning model. An integrated groundwater quality prediction framework was developed by Karunanidhi et al. [5] using WQI and supervised machine learning for assessing drinking water in a semi-arid river basin of South India. As well as this, Abu El-Magd et al. [6] used geospatial analysis, hydrochemical indices and machine learning algorithms with a view to improve groundwater quality assessment. Numerous studies have been conducted to analyze machine learning applications in a defined groundwater setting. Das and Das put forward a Water Quality Index with Objective Weights that, when combined with machine learning algorithms, can predict coastal groundwater quality. According to their findings, the reliability of groundwater classification improves with objective weighting as compared to conventional WQI computation. A similar study on the application of supervised machine learning algorithms for groundwater quality assessment and irrigation water quality prediction was done by Hussein et al. [9]. Through the classification, it was shown that ensemble-based classifiers have better performance than conventional classifiers. Recent research has also looked into combining geospatial technology with machine learning. The work talks about the investigation of sustainable water resource management in Shimla. It contains a chapter on project objectives, project description, groundwater simulation, result analysis, etc. Similarly, Md Isa et al. [10] applied spatial analysis techniques based on GIS for water quality assessment, which constitutes the second component of the text. Groundwater quality studies require hydrochemical investigations to examine. The Water Quality Index was used by Kushe et al. [11] to study post-monsoon groundwater quality in Maharashtra, India. Manna and Biswas [12] presented a review of WQI methodologies for studying drinking water. The studies conducted by Raj et al. [13–15] nowadays examined fluoride enrichment, health risk assessment, machine learning-assisted groundwater prediction and GIS techniques for groundwater contamination studies in South India. Researches show the periodical merger of hydrochemical study with intelligent computing methods. The application of machine learning technology has also been extended to water quality prediction. According to Sangwan and Bhardwaj [17], a supervised learning framework was developed for classifying water quality which shows improved prediction performance on both multiple water quality categories. Sarafaraz et al. [18] also aimed to evaluate a comparison of machine learning models with the QUAL2Kw model for predicting river water quality and concluded that data-driven learning approaches provide a better predictive ability for models of nonlinear environmental systems. The existing literature highlights a range of areas where limitations remain, despite progress. Multiple sources are aimed to forecast the continuous values of Water Quality Index rather than directly classifying the groundwater drinkability into practical categories. Many studies assess just one or two machine learning techniques, leaving little room for fair comparison. In addition, geographical diversity of hydrochemical characteristics may hinder the generalization of models. Accordingly, it still appears necessary to conduct an extensive comparative assessment of popular supervised machine learning algorithms using standardized hydrochemical datasets to establish reliable reference models for drinking water classification of groundwater.

Table 1. Comparative analysis of recent studies on groundwater quality assessment using machine learning

Ref.	Study	Dataset / Region	ML Techniques	Main Contribution	Limitation
[1]	Panigrahi et al. (2023)	CGWB, India	AI, ML	Groundwater drinkability prediction	Limited comparative analysis
[2]	Haggerty et al. (2023)	Review	ANN, SVM, RF, DT	Comprehensive ML review	No experimental validation
[3]	Huang et al. (2024)	Review	ML Models	Progress and challenges	Review only
[4]	Torres-Martínez et al. (2024)	Review	ML	Challenges and future directions	No benchmark comparison
[5]	Karunanidhi et al. (2025)	South India	ML + WQI	Integrated groundwater prediction	Single regional dataset
[6]	Abu El-Magd et al. (2023)	Egypt	ML + GIS	Hybrid geospatial framework	Limited classifier comparison
[7]	Das & Das (2024)	Coastal groundwater	ML + Objective WQI	Improved WQI weighting	Coastal region only
[8]	Sustainability (2024)	Groundwater	ML + GIS	Spatial groundwater prediction	Limited algorithm diversity

[9]	Hussein et al. (2024)	Groundwater	RF, SVM, XGBoost	Irrigation and groundwater prediction	Focus on irrigation index
[10]	Md Isa et al. (2024)	Lake water	GIS	Spatial water quality mapping	No ML comparison
[11]	Kushe et al. (2024)	Maharashtra	WQI	Groundwater assessment	Conventional WQI only
[12]	Manna & Biswas (2023)	Review	WQI	Review of WQI methods	No ML implementation
[13]	Raj et al. (2025)	South India	GIS	Fluoride risk assessment	No predictive ML
[14]	Raj et al. (2025)	South India	ML + GIS	Agricultural groundwater prediction	Limited classifiers
[15]	Meera Rajan et al. (2024)	South India	Hydrochemical analysis	Fluoride contamination assessment	No ML comparison
[17]	Sangwan & Bhardwaj (2024)	Water quality	ML	Water quality classification	General water dataset
[18]	Sarafaraz et al. (2024)	River water	ML + QUAL2Kw	Comparative predictive modelling	River water only
This Study	Proposed Work	CGWB, India (1345 samples)	LR, KNN, DT, SVM, AdaBoost, XGBoost	Comprehensive benchmark comparison for four-class groundwater drinkability prediction using hydrochemical features	Provides benchmark baseline for future explainable AI frameworks

3. EXPERIMENTAL STUDY

3.1 Water Quality Estimation Model (WQEM)

In this study, groundwater quality assessment is primarily based on the Water Quality Estimation Model (WQEM). The Water Quality Index generally known as WQI is widely used to assess the quality of water. It is a simple numeric measure of groundwater quality through the integration of more than one hydrochemical parameter into a single representative index. The Water Quality Index (WQI) simplifies complex water quality data for various usability. It has been used for environmental monitoring and water resource management. It is useful to assess the overall groundwater quality and the proper decision making process [12]. The result of WQI will depend on the choice of representative hydrochemical parameters that are able to represent the physicochemical properties of groundwater. At first, fourteen hydrochemical parameters published in Ground Water Year Book by Central Ground Water Board (CGWB), India were taken for the analysis. The most informative variables were determined using Exploratory data analysis (EDA), which includes statistical analysis, correlation analysis and outlier detection. From this evaluation, one parameter was excluded, the Carbonate (CO_3^{2-}), owing to the restricted contribution and variance in collected groundwater samples. Consequently, thirteen hydrochemical parameters were selected for computation of Water Quality Index (WQI) including pH, EC, TDS, TH, Alkalinity, Ca, Mg, Na, K, Cl, SO4, F and HCO3. The acceptable limits for these hydrochemical parameters were followed the World Health Organization's (WHO) Guidelines for Drinking-water Quality and Bureau of Indian Standards (BIS IS 10500:2012, Revised 2015). The computation of Water Quality Index and classification for drinking of groundwater of safe site. The selected hydrochemical parameters and their permissible values are presented in Table 3.1. Water Quality Index can be set as an indicator for assessing groundwater quality; based on it, groundwater samples are categorized into four classes (Excellent, Good, Poor, and Bad) of drinkability which are used for supervised machine learning-based classification.

Table 2. Bhopal Permissible Value for Parameters Used in Calculating Water Quality Index(WQI)

S. No.	Parameters	Permissible Values (μg)
1	PO_4	5 - 6.5
2	P_H	6.5 - 8.9
3	Sodium	210
4	Electrical Conductivity(EC)	1100
5	Potassium	13
6	Total Dissolved Solids(TDS)	510
7	Carbonate	00
8	Total Hardness	315
9	Bicarbonate	365
10	Alkalinity	210
11	Chloride	265
12	Calcium	85
13	Sulphate	210
14	Magnesium	60
15	Fluoride	1.8

3.2. Water Quality Index Model

A widely accepted measure of groundwater quality is the Water Quality Index (WQI). It provides a numerical value that integrates many hydrochemical parameters to represent groundwater quality. It makes assessment of ground water for drinking easy with a simple yet comprehensive representation of water quality. The WQI system has been widely used in many environmental and hydro-geological studies. The reason is it allow for an easy comparison of water quality in diverse places of the world. It also reduces the complexity behind the many physicochemical parameters [12]. In this work, WQI was computed using thirteen selected hydrochemical parameters such as pH, EC, TDS, TH, Alkalinity, Ca, Mg, Na, K, Cl, SO4, F and HCO3. Selection of these parameters were done after exploratory data analysis but are in agreement with the acceptable drinking water quality standards as stipulated by the World Health Organization (WHO) and Bureau of Indian Standards (BIS). The Water Quality Index is a weighted average of the quality ratings of all selected hydrochemical parameters and is defined as:

$$\mathrm{WQI} = \frac{\sum_{i=1}^{N} q_i\, w_i}{\sum_{i=1}^{N} w_i} \qquad (1)$$

In this equation, symbol 'N' signifies the total number of hydrochemical parameters. This was used for calculating WQI (in this study, it is, $N = 13$). q_i is the quality rating of the i^{th} parameter. w_i is the unit weight of the ith parameter. The quality rating (q_i) for each hydrochemical parameter is worked out as.

$$q_i = 100\left(\frac{V_i - V_{\text{ideal}}}{S_i - V_{\text{ideal}}}\right) \quad (2)$$

where V_i is the experimentally measured concentration of the i^{th} hydrochemical parameter, S_i represents the corresponding permissible standard value specified by WHO/BIS, and V_{ideal} denotes the ideal concentration of that parameter in drinking water. For most hydrochemical parameters, the ideal value is assumed to be 0**,** whereas for pH, the ideal value is 7.0. To account for the relative importance of each hydrochemical parameter, a unit weight (w_i) is assigned inversely proportional to its permissible concentration. The unit weight is calculated as:

$$w_i = \frac{K}{S_i} \quad (3)$$

where K is a proportionality constant determined by

$$K = \frac{1}{\sum_{i=1}^{N}\left(\frac{1}{S_i}\right)} \quad (4)$$

The computed WQI provides an overall measurement of groundwater quality in a quantitative way. Lower values of WQI mean better quality of water. Increasing values mean increasing level of contamination. The computed WQI values are then assigned to predefined categories of groundwater quality, which are further used for multi-class drinkability prediction using supervised machine-learning algorithms.

3.2. Water Quality Class Model

The Water Quality Class (WQC) model makes use of the Water Quality Index (WQI) to turn continuous values into discrete quality classes of groundwater so as to make it easier to understand the groundwater quality for drinking purposes. Though the WQI shows the overall quality of water, its classification into different quality classes aids water resource managers in decision-making as well as helps in developing supervised machine models. The groundwater samples' WQI values are computed and based on that they are classified into five different classes according to WHO drinking water quality guidelines and widely used WQI classification. Groundwater of higher quality will have lower WQI value and vice versa. Higher value indicates high pollution and less suitable for drinking. According to classification, groundwater is classified as:

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

The traditional WQI classification defines five categories of groundwater quality but the current work aims at automated prediction of groundwater drinkability with the use of supervised machine learning. To ensure convergence of the simulation classrooms and improve classification performance, the groundwater samples are classified into families of four so-called representative drinkability classes, namely Excellent, Good, Poor, and Bad. The Very Poor and Undrinkable classes are combined and referred to as Bad class. This simplified classification results in a more balanced multi-class learning problem whilst still preserving the practical interpretation of groundwater quality for drinking purposes.

3.3. Problem Formulation

The assessment of groundwater quality could be framed as a supervised multi-class classification problem, optimally predicting the drinkability class of a groundwater sample based on its hydrochemical properties. Each groundwater sample has a number of physicochemical parameters recorded. While the Water Quality Index for each sample is computed using the method discussed in section 3.2 and mapped to one of the groundwater quality classes defined in section 3.3. Let the groundwater dataset is represented as.

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$$

defined as $x_i = [x_1, x_2, \dots, x_n]$ the feature vector built by 13 hydrochemical parameters, y_i the corresponding groundwater quality class, and M the total number of groundwater samples. The hydrochemical feature vector incorporates pH, EC, TDS, TH, Alkalinity, Ca, Mg, Na, K, Cl, SO_4^{2-}, F^-, and HCO_3^- in the study. We want to learn a classification function.

$$f : x \rightarrow y \qquad (6)$$

that accurately predicts the groundwater drinkability class:

$y \in \{\text{Excellent}, \text{Good}, \text{Poor}, \text{Bad}\}$ from the corresponding hydrochemical measurements. To achieve this objective, a comprehensive machine learning framework is adopted. Initially, groundwater samples undergo exploratory data analysis, missing value handling, outlier detection, correlation analysis, and feature normalization to improve data quality and model robustness. The processed dataset is subsequently used to train and evaluate six supervised machine learning algorithms, namely Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), AdaBoost, and Extreme Gradient Boosting (XGBoost). Their predictive performances are compared using standard evaluation metrics, including Accuracy, Precision, Recall, and F1-score, to identify the most suitable model for groundwater quality assessment and drinkability prediction. Figure 1 illustrates the overall workflow of the proposed machine learning-based groundwater quality assessment framework.

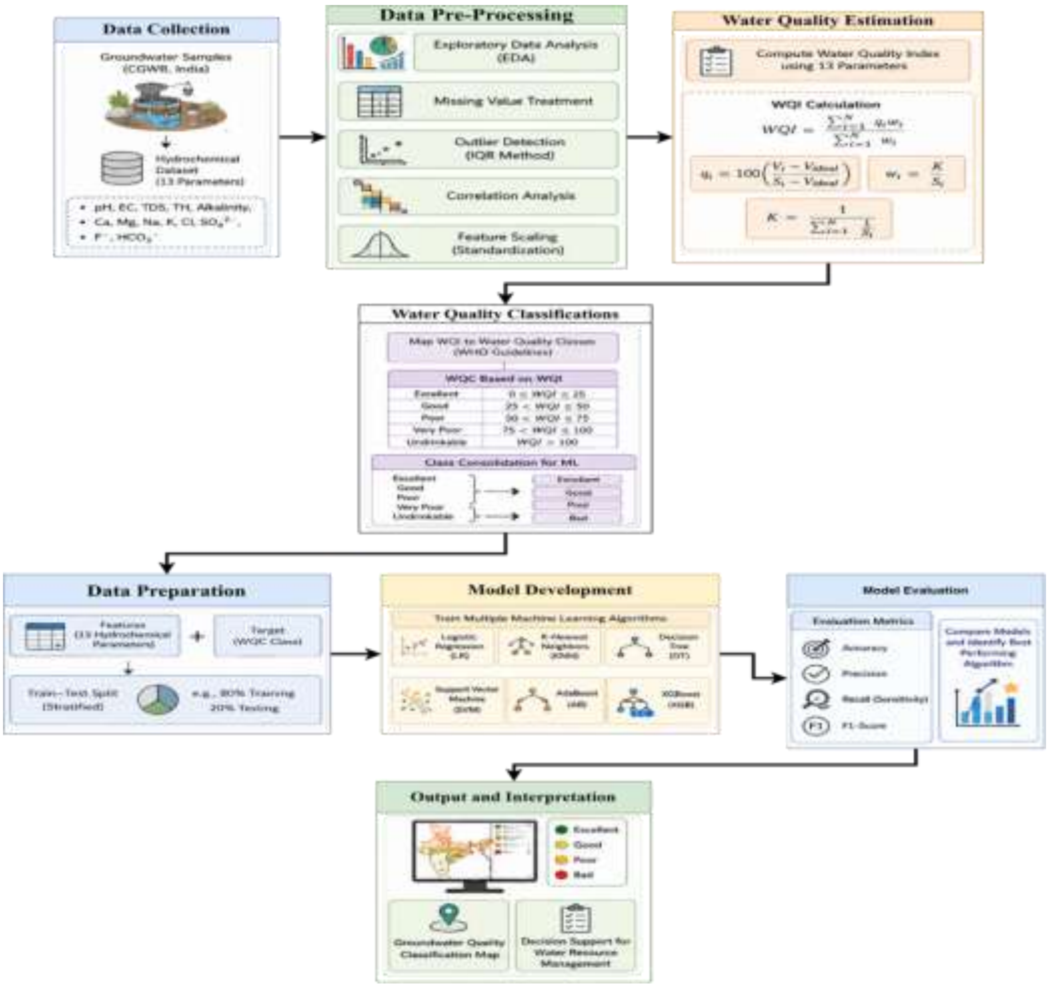

Figure 1. workflow of the proposed machine learning-based groundwater quality assessment framework

3.4. Proposed flow diagram

The overall workflow of the proposed machine learning-based groundwater quality assessment framework is illustrated in Figure 1. The framework is designed to automate groundwater quality assessment and drinkability prediction using hydrochemical characteristics through a systematic sequence of data acquisition, preprocessing, Water Quality Index (WQI) computation, multi-class classification, machine learning model development, and performance evaluation. The proposed workflow consists of five major phases, as described below. In the first phase, groundwater samples are collected from the Central Ground Water Board (CGWB), India, together with their corresponding hydrochemical measurements. The collected dataset contains thirteen significant hydrochemical parameters, namely pH, Electrical Conductivity (EC), Total Dissolved Solids (TDS), Total Hardness (TH), Alkalinity, Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Chloride (Cl), Sulphate (SO_4^{2-}), Fluoride (F^-), and Bicarbonate (HCO_3^-), which serve as the input features for groundwater quality assessment. The second phase focuses on data preprocessing to improve data quality and enhance the robustness of the predictive models. Initially, exploratory data analysis (EDA) is performed to understand the statistical characteristics of the groundwater samples. Subsequently, missing values, inconsistent observations, and abnormal measurements are identified and handled through appropriate data cleansing procedures. Correlation analysis is then carried out to investigate the relationships among hydrochemical parameters and their influence on the Water Quality Index. Finally, feature normalization is applied to transform the hydrochemical variables into a common numerical scale, thereby improving the convergence and prediction performance of machine learning algorithms. In the third phase, the Water Quality Index (WQI) is computed using the selected hydrochemical parameters according to the standard weighted arithmetic index method. Based on the computed WQI values, groundwater samples are categorized into predefined drinkability classes following the Water Quality Class (WQC) model. For the purpose of supervised learning, the groundwater samples are grouped into four representative categories: Excellent, Good, Poor, and Bad, thereby formulating the groundwater quality assessment problem as a multi-class classification task. The fourth phase involves the development of predictive machine learning models. The processed dataset is divided into training and testing subsets using an appropriate train–test splitting strategy. Six widely adopted supervised machine learning algorithms, namely Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), AdaBoost, and Extreme Gradient Boosting (XGBoost), are trained using the hydrochemical features. The objective of this phase is to identify the most suitable machine learning algorithm for groundwater quality assessment and drinkability prediction. In the final phase, the trained models are evaluated using standard classification performance metrics, including Accuracy, Precision, Recall, and F1-score. The comparative analysis enables objective evaluation of the predictive capabilities of each algorithm and facilitates the identification of the best-performing model for groundwater quality prediction. The selected model can subsequently support rapid groundwater quality assessment and assist environmental agencies and water resource managers in making informed decisions regarding drinking water safety. Figure 1 presents the complete workflow of the proposed machine learning-based groundwater quality assessment framework, illustrating the sequential process from groundwater data acquisition to automated drinkability prediction and performance evaluation.

3.5. Proposed Machine Learning Strategy

The machine learning approach proposed in this study follows a defined workflow as shown in Figure 1, which combines groundwater quality assessment with supervised machine learning for automated drinkability prediction. The framework includes systematic sequences of data acquisition and pre-processing, Water Quality Index (WQI) computation, groundwater quality classification, model development, and performance evaluation. Groundwater samples are primarily collected from the Central Ground Water Board (CGWB) database, and relevant hydrochemical parameters are extracted. The data will then be carried out exploratory data analysis followed by cleansing, correlation analysis, outlier detection and normalisation of features to enhance the data quality for training a reliable model. After performing data preprocessing, Water Quality Index (WQI) is determined using the weighted arithmetic index method. The surveyed groundwater samples were designated into classifications of drinkability based on the computed water quality index (WQI) values. The treated dataset is divided into training and testing subsets. Six supervised machine learning algorithms, Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), AdaBoost and Extreme Gradient Boosting (XGBoost) are trained to predict groundwater quality classes. In the end, standard model evaluation metrics such as Accuracy, Precision, Recall, F1-score, etc. are used to find the best model (algorithm) for groundwater quality assessment and drinkability prediction.

3.6. Study Area and Dataset Description

During the study, groundwater quality data for the state of Madhya Pradesh, India – one of the largest states in central India where groundwater is used as a drinking water source by urban and rural people – was collected. Considering the natural geological processes and human interventions, groundwater quality exhibits appreciable variation in space. The study area

Natural Resources for
Human Health
eISSN: 2583-1194

consists of various hydrogeological formations. The hydrogeological map including the position of Bhopal, the Madhya Pradesh capital, for the study area is (figure 2(a)).Quality of groundwater data was acquired from the Central Ground Water Board (CGWB), Ministry of Jal Shakti, Government of India. The Central Ground Water Board operates a detailed network of National Hydrograph Network Stations (NHNS), consisting of observation wells, dug wells and piezometers across the districts of Madhya Pradesh. Water samples are collected on a regular basis and hydro-chemical analysis are being carried out at these monitoring stations. The monitoring stations shall be distributed carefully across the area under study. Based on the concentrations of the major hydrochemical constituents concentrations of groundwater samples from the NHNS monitoring stations collected annually and analyzed in accredited regional laboratories. The Ground Water Year Book is published after analyzing the data. The Ground Water Year Book is an authoritative source of groundwater quality in India. In this research paper the 2018 – 2019 year Ground Water Year Book quality of groundwater records were extracted and converted into a structured comma-separated values (CSV) dataset for further analysis using machine learning techniques. This study provides a dataset indicating groundwater of the area is mineralized as well as of good quality because total hardness is 57 and sulphate is 92 mg/l along with other essential characteristics like pH, E.C., T.D.S. and others. The groundwater is deep and has a good yield. The parameters serve as input features for Water Quality Index calculations and the subsequent prediction of groundwater drinkability using supervised machine learning models.

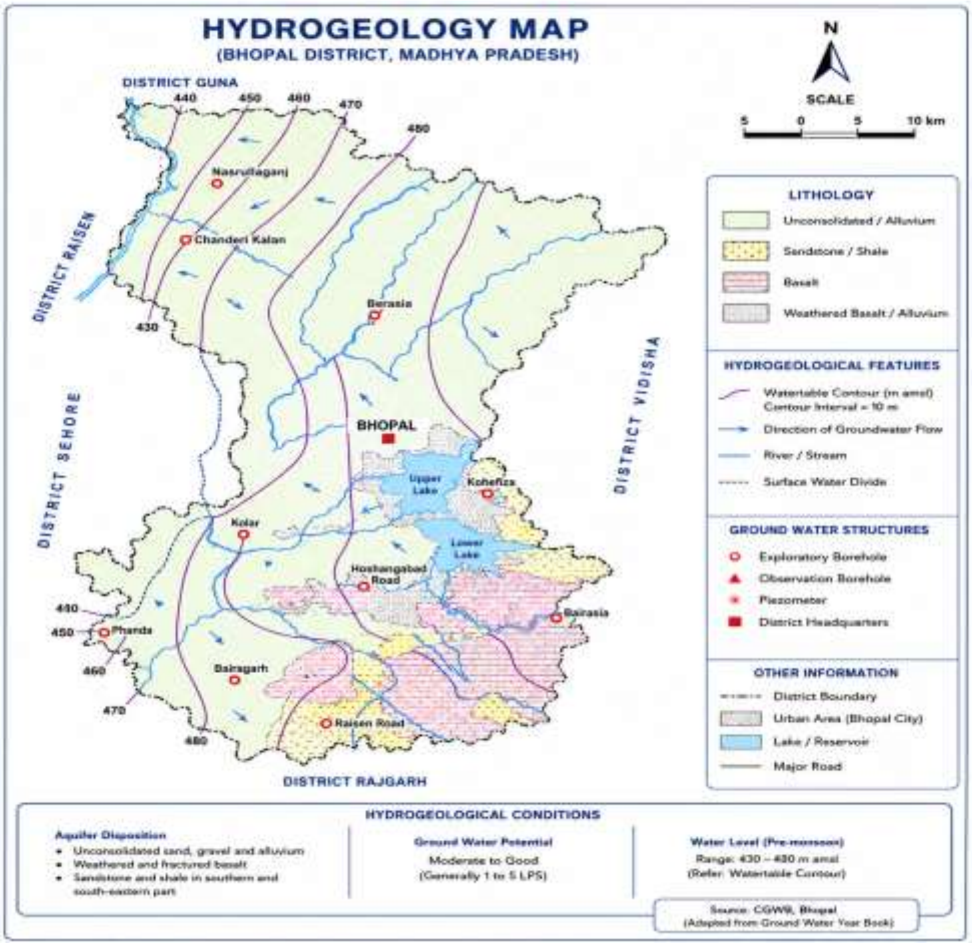

Figure 2. (a) Hydrogeological map of Madhya Pradesh showing the study area

Figure 2(b). Distribution of National Hydrograph Network Stations (NHNS) used for groundwater sample collection

3.7. Exploratory Data Analysis (EDA)

Before the development of machine learning predictive models for groundwater data sets, it is important to do exploratory data analysis (EDA) to have a clear understanding of groundwater data set. The main objective of EDA is to examine the statistical properties of the collected hydrochemical parameters and identify missing or erroneous observations, assess outliers, distribution of features and relationships among features. Increasing the quality of data and obtaining meaningful information helps in the important preprocessing, feature engineering and modelling analysis. The groundwater dataset comprises 1,345 groundwater samples from different districts of Madhya Pradesh, India, used in this study. Every sample consists of fourteen hydrochemical parameters obtained from the Ground Water Year Book (2018-2019) published by the Central Ground Water Board (CGWB). Each hydrochemical parameter was subjected to descriptive statistical analysis to present their characteristics. As shown in Tables 3-5, which include the Count (number of observations), Mean, Std (the standard deviation), Min (the minimum value), Max (the maximum value), and percentiles (25, 50, and 75). By looking at these measures we get the initial impression about the distribution, variability and central tendency of a parameter. The descriptive statistics show great variability among the hydrochemical parameters indicating that groundwater quality varies greatly in the study area. For example, the median (50th percentile) value of pH is around 7.9, which means that most groundwater samples are neutral to slightly alkaline. Like wise, Electrical Conductivity (EC) was moderately variable, with nearly half of ground water samples having EC less than 550 µS/cm, indicating noticeable difference in dissolved ionic content among sampling sites.

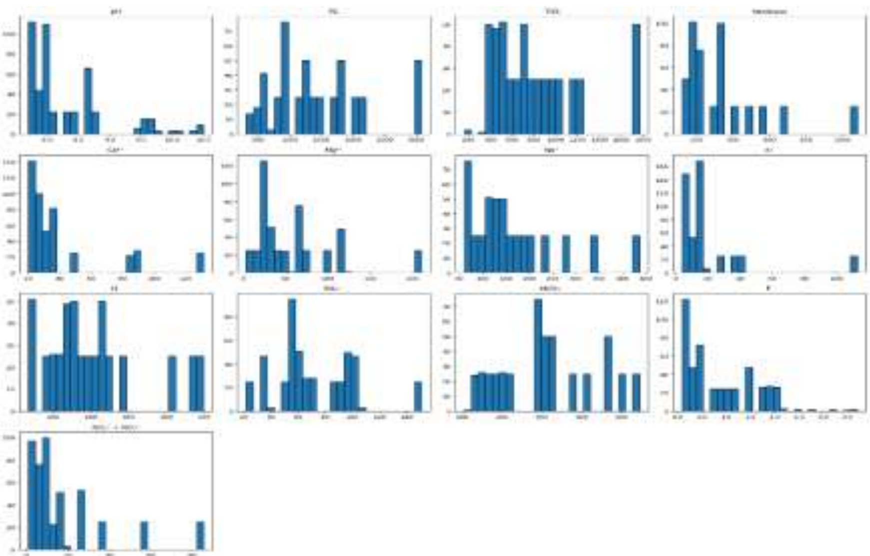

Figure 3 (b). Distribution of Hydrochemical Psrameters

An exploratory analysis of the Carbonate (CO_3^{2-}) parameter reveals it to have a very low variation with a 50th percentile (median) and 75th percentile equal to zero. Most groundwater samples were observed to show that there was little or no carbonate in them and therefore, does not add much to the quality dissimilarity. As a result, the Carbonate is removed from the analyses to avoid redundancy in the prediction. The other thirteen hydrochemical parameters were retained as an input feature for Water Quality Index computation and for ML based groundwater quality assessment. Subsequent EDA was done to identify strange observations that might impact the model negatively. The parameters that exhibit values significantly outside their expected statistical ranges were assessed as possible outliers. To ensure accuracy, all non-zero values of the Carbonate parameter were considered anomalous and excluded before the preprocessing stage. Through the identification and treatment of such outliers, the reliability of the dataset gets improved, and the robustness of the subsequent ML models gets better. Exploratory data analysis confirm that the groundwater dataset post suitable preprocessing is fit for predictive modelling. The statistical results achieved from EDA provide a basis for doing correlation analysis, feature normalization and comparative performance evaluation of supervised machine learning algorithms in the following sections.

Table 3: Descriptive statistics of major hydrochemical parameters (Part I)

Metrics	Alkalinity	TDS	EC	PH	TH
Count	1243	1243	1243	1243	1243
Mean	180.500	360.090	699.5	7.89	217.020
Standard	105.940	283.981	540.830	.405	161.790
Minimum	16	32	7.25	6.51	25
25%	107	190	371	7.61	128
50%	160	282	565	8.3	190
70%	231	451	923	8.26	271
Maximum	770	2790	5791	8.90	1950

Table 4: Descriptive statistics of major hydrochemical parameters (Part II)

Metrics	Ca	Potassium	Na	Mg
Count	1243	1243	1243	1243
Mean	45.351	14.134	50.159	26.516
Standard	32.315	31.157	62.055	26.653
Minimum	0	0	0	-5
25%	29	2.2	19	13
50%	41	4.3	35	21
70%	56	11.2	70	36
Maximum	501	340	826	349

Table 5: Descriptive statistics of major hydrochemical parameters (Part III**)**

Metrics	Chloride	Fluoride	Bi-Carbonate	Sulphate
Count	1243	1243	1243	1243

Mean	45.351	14.134	50.159	26.516
Standard	32.315	31.157	62.055	26.653
Minimum	0	0	0	-5
25%	29	2.2	19	13
50%	41	4.3	35	21
70%	56	11.2	70	36
Maximum	501	340	826	349

3.8.1 Data Cleansing and Outlier Detection

Outlier detection and data cleansing are performed to make the groundwater dataset reliable and consistent to develop machine learning models. Hydrochemical datasets may experience data loss, inconsistent observations, and outliers caused either by measurement error, sampling variability, or the hydrogeochemical processes. Improper assessment of such observables could narrowly skew the statistics and can worsen the predictions of machine learning algorithms. Consequently, the dataset will be screened for potential outliers before further analysis. The boxplot-based analysis in the study is used to show the distribution and dispersion of individual hydrochemical parameters. It helps to identify the observations that lie significantly above the typical range. Special care is given to some parameters like Fluoride (F^{-}) and pH which having extreme values may be important for drinking-water quality. Instead of viewing every statistically extreme observation as an error by default, the identified values are checked with reference to adopted permissible drinking-water limits in the study. This distinction is important because extreme concentrations may be contaminants in groundwater or hydrogeochemical variability and not measurement or input errors. The same screening process is applied for the rest of the hydrochemical parameters to check for outliers and inconsistencies. Values that are clearly invalid or inconsistent are dealt with in the data cleansing process. Conversely, extreme concentrations that do represent actual groundwater quality trends are retained. This method helps in avoiding loss of important information while also reducing the influence of erroneous characteristic measurements on subsequent water quality index and machine learning-based classification. Figure 3(a) represents the boxplot distributions of the selected hydrochemical parameters. The visualization shows that there are differences in ranges of the specific features and also indicates the parameterswhich exhibit relatively higher dispersion with extreme possible observations. The dataset cleaned after this process is later utilized for correlation analysis, normalization, WQI and machine learning model formulation.

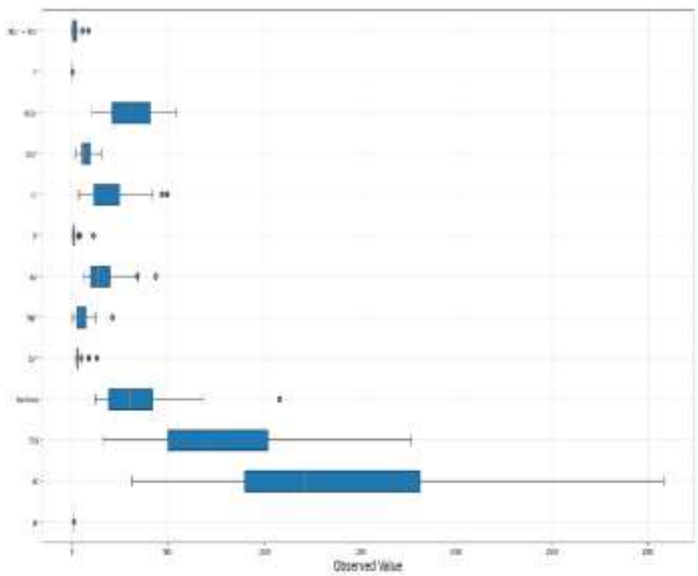

Figure 3(a). Boxplot distribution of hydrochemical parameters for data cleansing and potential outlier identification

3.8.2 Correlation Analysis of Hydrochemical Parameters

A correlation analysis was done to investigate the degree of association and direction of association among the hydrochemical parameters of groundwater. This is because groundwater chemistry is controlled in space and time by interacting hydrogeochemical processes so that many measured parameters will contain related or overlapping information. As a result, correlation analysis can assist in identifying possible relations among input variables, strongly associated features, and more insight into the physicochemical parameters of groundwater in the study area. The study uses a method of calculation referred to as the Pearson correlation coefficient (r) which measures the linear relationship between two variables. The Pearson correlation coefficient is given by ϱx,y for two variabless.

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (7)$$

In this context, X_i and Y_i refer to the observations of X and Y variables respectively, while $\bar{X}$ and $\bar{Y}$ refer to the mean values of X and Y variables respectively and n is the number of groundwater samples. The coefficient varies from -1 to +1, whereby values close to +1 indicate a strong positive linear relationship, values close to -1 indicate a strong negative relationship, and zero indicates a weak linear relationship. The correlation findings presented in the paper show several interesting associations. EC and TDS show a very high positive correlation (r =0.986). It means that increased concentration of dissolved ionic constituents is closely reflected by increased electrical conductivity. Total Hardness (TH) also had stronger associations with Magnesium (r =0.901), electrical conductivity (EC) (r =0.852), total dissolved solids (TDS) (r =0.826) and Calcium (r =0.797). The relationships established align with the dataset structure that ascribe hardness in groundwater mainly to Ca^{2+} and Mg^{2+}. The correlation matrix also suggests that among the variables included in that matrix, Alkalinity (r = 0.641), TDS (r = 0.601) and EC (r = 0.581) show relatively high positive correlation with WQI. The relationship between pH and WQI is relatively moderate (r =0.400), on the other hand. The studied groundwater samples' calculated WQI mainly depends on the variations in the dissolved minerals and alkalinity. To visualize these interrelationships in a more straightforward way, their complete correlation structure should be visualized through a correlation heatmap (figure 4), while the numerical Pearson correlation coefficients can be kept in the corresponding table. The joint graphical and numerical analysis allows us to identify feature dependencies in a systematic way which further serves in the preprocessing (stage one) and machine learning (stage two) steps.

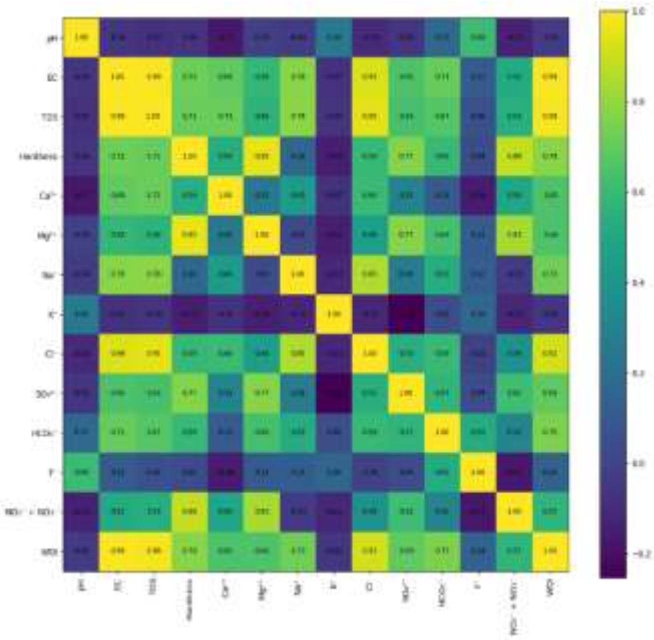

Figure 4. Pearson correlation heatmap illustrating the relationships among hydrochemical parameters in the Madhya Pradesh groundwater dataset

Table 6. Observation from Co-relation Analysis of Different Water Parameters

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

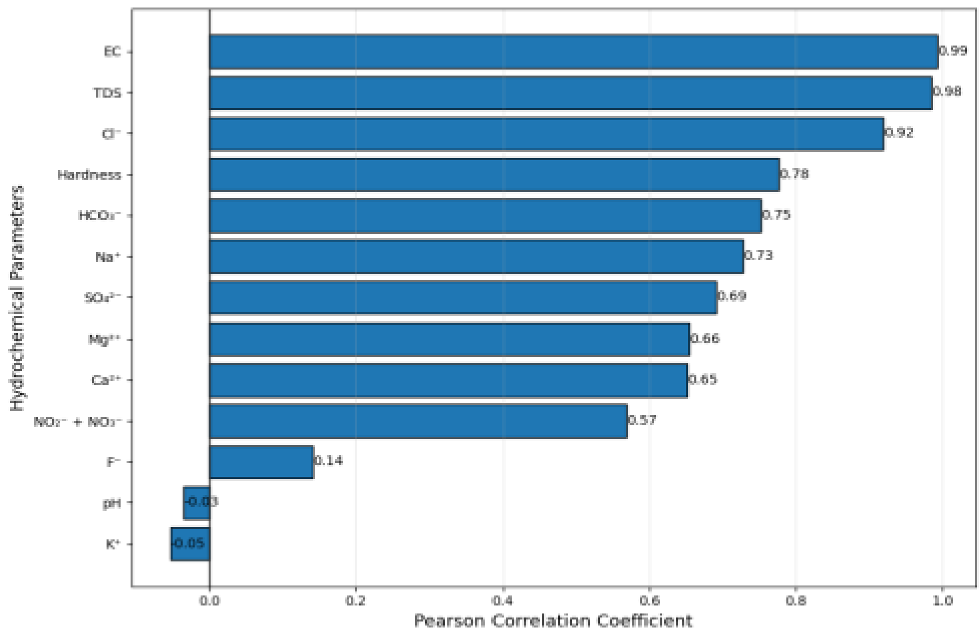

Figure 5. Co-Relation Analysis of Water Quality Index with other Water Parameters

It is observed from Figure 5. that all parameters are positively correlated with WQI.

3.8.3 Data Preprocessing

The first important stage of the groundwater quality assessment framework proposed is data preprocessing as the input data quality and consistent application of machine learning algorithms affect prediction performance of the proposed framework. Hydrochemical parameters in raw groundwater datasets might possess different units/numerical range; this bias model training and reduces the prediction performance of the models. Hence, suitable preprocessing techniques are used to render the dataset in a standardized format for supervised machine learning. After identifying outliers and cleaning up the data, as well as performing correlation analysis, the selected hydrochemical parameters are normalized so that any scale does not affect their learning. Normalizing features removes the influence of variations in measurement scales while retaining variation among groundwater samples. As a result, machine learning models become more stable and converge more quickly during training. The results are also more reliable during classification. In this study, the Min–Max normalization technique was used to transfer each hydrochemical parameter into the interval [0, 1]. In contrast to standardization methods, which transform the distribution of data, Min–Max normalization achieves the same effect while also limiting the influence of features with greater numerical ranges during learning. This method is particularly useful for groundwater data sets containing significantly different measurement units for hydrochemical parameters. The value of each feature is computed as normalized.

$$x_{sc} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (8)$$

where x denotes the original value of a hydrochemical parameter, $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the corresponding feature in the dataset, respectively, and x_{sc} is the normalized feature value. The resulting normalized dataset is subsequently used for Water Quality Index-based classification and machine learning model development.

3.8.4 Machine Learning Models and Performance Evaluation

After preprocessing, the normalized groundwater dataset is used to build supervised machine learning models for groundwater quality assessment and drinking prediction. The aim of this is to find out the best classification algorithm for the prediction of groundwater quality classes. The split of the data is done in two sets. The predictive models are constructed using the training dataset and tested on the testing dataset to evaluate their capacity for generalization to new groundwater samples. We analyze six supervised machine learning algorithms in this study; they are Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), AdaBoost and Extreme Gradient Boosting (XGBoost) . The algorithms were picked for their diversity of learning paradigms, which included a linear classifier, distance-based learning, tree-based learning, kernel-based classification, and ensemble learning. Before the training of each machine learning model, an appropriate set of hyperparameters initializes the model. Hyperparameters control the learning behaviour of the respective algorithm and have a huge impact on predictive performance. The default hyperparameter settings provided by the Scikit-learn and XGBoost library will be used unless specified otherwise. However, algorithm-specific parameters will be adjusted if necessary to improve classification performance. The hyperparameter configuration details for the evaluated models is mentioned in Table 6. After training the models, each classifier is evaluated for predictive performance on an independent testing data set. The accuracy, precision, recall, F1-score etc. standard multi-class classification metrics are comprehensively compared. Furthermore, confusion matrices are used to evaluate the prediction ability of each model on each class and identify misclassifications. The comparative evaluation allows suitable selection of ML Algorithm for groundwater quality evaluation and drinkability prediction in the study area.

3.9 Machine Learning Models

This study screened six supervised machine learning algorithms to find the best predictive model for groundwater quality assessment and drinkability prediction. These algorithms reflect different kinds of learning models such as linear models, distance-based classifiers, tree-based learning and many other. Each of the models has been trained on a normalized hydrochemical dataset and evaluated in a similar experimental setting for a fair comparison. The models are evaluated for prediction accuracy using standard multi-class classification metrics such as Accuracy, Precision, Recall, and F1-score.

3.9.1 Logistic Regression (LR)

Logistic Regression (LR) is one of the popular supervised learning algorithms for classifying problems. LR utilizes the logistic (sigmoid) function and, unlike linear regression, it estimates the probability of a sample belonging to a class. In this study, the multinomial logistic regression variant is adopted, since the groundwater drinkability prediction involves multiple categories of quality. When you have the hydrochemical feature vector x, the probability of assigning the sample to class k is given by the Softmax function.

$$P(y = k \mid x) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1}^{K} \exp\left(w_j^T x + b_j\right)} \tag{9}$$

where K is the total number of classes of groundwater quality, $w_k$is the weight vector for class k, and b_k is its bias. Logistic Regression is a computationally efficient and easy to interpret model. Most importantly, a Logistic Regression works well enough when classes are linearly separable. In the present paper, we use it as a baseline classifier to evaluate the performance of more complex machine learning models.

3.9.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple supervised machine learning algorithm that classifies data based on the distance of nearby data points in the feature space. The Euclidean distance, which is defined as, is used to measure similarity between groundwater samples.

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - x_{jm})^2} \tag{10}$$

Where n summarizes the hydrochemical parameters. The K nearer neighbours of a new groundwater sample are found using the algorithm and the class which has been voted most is assigned. Since KNN performs distance calculations, it is necessary to have features normalized to ensure that non-dominating variables do not have larger numerical range. KNN is a suitable modeling mechanism for nonlinear decision boundaries and thus makes a good benchmark for groundwater quality classification.

3.9.3 Decision Tree (DT)

Decision Tree (DT) is a supervised learning algorithm that separates the features from the target by recursively splitting the feature space into rules or conditions based on the input features. While building the model, the algorithm selects the hydrochemical parameter which has maximum information gain or impurity reduction and splits the dataset along child nodes. This recursive overall always terminates on meeting some general stopping criterion. Gini index assigns a numerical value to impurity of node.

$$G = 1 - \sum_{i=1}^{K} p_i^2 \tag{11}$$

where p_i is the probability of class i inside the node. Decision Trees are develop complex nonlinear relationships which are also usable in hierarchical structure. These models are suited for the assessment hydrochemical parameters interaction to groundwater quality.

3.9.4 Support Vector Machine (SVM)

Support vector machine is a supervised learning algorithm. SVM in machine learning is used for constructing a hyperplane. When quality classes of groundwater cannot be separated by a linearly separable function, kernel functions are used to convert the original feature space to higher dimensional space where separation possible. The SVM decision function is denoted by.

$$f(\boldsymbol{x}) = w^T \phi(x) + b \tag{12}$$

The quantity where w represents the weight vector and b represents the bias term is expressed as $\phi(x)$. In the study, various kernel functions were investigated during preliminary experiments; and the kernel that produced the best classification results was used to predict drinkability groundwater. The SVM or support vector machine is effective for high-dimensional datasets and gotten strong predictive performance on environmental classification problems.

3.9.5 AdaBoost

To construct a strong classifier by combining weak classifiers, AdaBoost (Adaptive Boosting) is an ensemble learning algorithm. Throughout each cycle, the algorithm attaches more importance to the wrongly classified groundwater samples by increasing their weights. Hence, it forces the following weak learners to focus on the difficult observations. The final prediction is reached via a weighted sum of classifiers.

$$F(x) = \sum_{t=1}^{T} \alpha_t \, h_t(x) \qquad (13)$$

In (3), $h_t(x)$ represents the prediction of the t^{th} weak learner, α_t is its associated weight, and T is the number of boosting iterations. AdaBoost is known to reduce the bias of supervised classifiers. Typically it performs better than individual decision trees in classification. It can be applied in groundwater quality prediction with heterogeneous hydrochemical characteristics due to its ensemble learning capability.

3.9.6 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting, popularly known as XGBoost, is a specialized ensemble learning algorithm. The technique builds predictive models in sequence, with every new tree minimizing an error metric of the previous tree. XGBoost uses second-order gradients and regularization to construct trees in parallel, improving prediction accuracy and computation speed. XGBoost optimizes the following objective function.

$$\mathcal{L} = \sum_{i=1}^{N} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\,(f_k) \qquad (14)$$

In this expression, $l(\cdot)$ is the training loss function, $\hat{y}_i$are the predicted outputs, and $\Omega(f_k)$ is the regularization term that governs model complexity. Due to robustness towards overfitting with respect to lesser irrelevant feature interactions and higher predictive power, XGBoost becomes one of the most effective machine learning algorithms for environmental data analysis. This study will assess the suitability of XGBoost in comparison to other supervised learning algorithms for groundwater quality assessment and drinkability prediction.

3.10 Performance Evaluation Metrics

The assessment of the machine learning models is made via commonly used multi-class classification metrics. These evaluation measures quantify the ability of each classifier to correctly classify the groundwater drinkability classes and allows for a detailed comparison of the competing algorithms. The evaluation of performance is conducted utilizing the confusion matrix, an application from which the metrics Accuracy, Precision, Recall, and F1-score are derived. All the above measures on the whole provide helpful complementary information regarding the accuracy, reliability, and robustness of the analysis groundwater quality prediction models.

3.10.1 Confusion Matrix

The confusion matrix is a table that summarizes the classification results based on comparing the predicted groundwater quality classes to the actual classes. It gives all the groundwater samples that are classified correctly and incorrectly which is helpful in calculating other evaluation matrix. In a multi-class classification problem, confusion matrix has four basic elements:

- **True Positive (TP):** Number of groundwater samples correctly classified into their actual quality class.
- **True Negative (TN):** Number of groundwater samples correctly identified as not belonging to a particular class.
- **False Positive (FP):** Number of groundwater samples incorrectly assigned to a quality class.

- **False Negative (FN):** Number of groundwater samples incorrectly classified as another quality category.

A confusion matrix helps analyze how well a model's classification works by seeing what kinds of predictions are wrong for each class. It also helps evaluate a model's ability to discriminate between classes.

3.10.2 Accuracy

Accuracy measures the overall proportion of correctly classified groundwater samples among all evaluated samples. It provides a general indication of the predictive capability of a classification model and is one of the most widely used performance measures. The classification accuracy is computed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (15)$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Higher accuracy indicates better overall predictive performance. However, accuracy alone may not adequately reflect model performance when the class distribution is imbalanced.

3.10.3 Precision

Precision refers to the percent of groundwater samples predicted to belong to a particular quality class that are classified correctly. It assesses how reliable predictions are positive and is a useful metric when the goal is to minimize false-positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (16)$$

A higher value of precision indicates that the classifier generates fewer false alarms with reasonably accurate prediction of groundwater quality.

3.10.4 Recall

Recall, or Sensitivity also known as the True Positive Rate is the classifier's ability to identify as groundwater samples which belong to a particular class of quality. It indicates the ratio of true positives that are correctly predicted by the model. The calculation of recall is done as:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (17)$$

Higher recall means that fewer groundwater samples are misclassified into one drinkability category or another.

3.10.5 F1-Score

The f1-score measures average precision in proportion to recall. In other words, it is derived using the harmonic mean. This offers an even-handed evaluation of how well classification works. In addition, it is especially useful for datasets with imbalanced class distributions. The F1 score is given by:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (18)$$

A classifier with high precision and high recall leads to a higher F1-score which assesses the balance between the false positives and false negatives.

4. RESULTS SUMMARY & DISCUSSION

4.1 Comparative Performance Analysis of Machine Learning Models

Using the testing dataset, the predictive performances of the six supervised machine learning algorithms were compared. The performance was evaluated on standard multi-class classification metrics such as Accuracy, Precision, Recall, and F1–score using all models. It was done to obtain the best suitable model for groundwater quality assessment and drinkability. The overall classification accuracy obtained by machine learning models is shown in Figure 8. Accuracy indicates how many of the groundwater samples which are accurately classified into their drinkability classes. The comparative study allows one to impartially assess the predictive power of different algorithms and to also understand which classifier is best for the groundwater quality data set. The model which gives the highest accuracy indicates better generalization capacity thus selected as the best model for groundwater drinkability prediction. The six machine learning algorithms were evaluated on precision across four classes: Excellent, Good, Poor and Bad. The results can be seen in Figures 9. Precision assesses the dependability of the predicted classification of groundwater quality classes is determined by the ratio of True positives to all predicted positives. A higher precision value indicates that such a classifier makes predictions that do not misclassify groundwater quality. Likewise, Figure 10 compares the F1 score of the machine learning model among the four classes of groundwater quality. F1-score effectively combines both measures into one through their harmonic mean, giving us an overall assessment which weighs precision as well as recall. This metric is a useful evaluation metric that takes false-positive and false-negative errors into account in a single score, and is particularly useful in the case of multi-class classification problems. As depicted in Figure 11, class-wise Recall values for evaluated models exist in the machine learning environment with the help of the Recall value, which correctly identifies the groundwater samples belonging to the drinkability class. A higher Recall value implies that the groundwater samples classified into other quality classes are less. Therefore, it enhances groundwater quality assessment analysis. The average values of Accuracy, Precision, Recall, and F1 score of the evaluated algorithms are summarised in Table 7 for an overall comparison. The comparative results reveal which machine learning model will be the best fit for groundwater quality assessment based on predictive accuracy and classification strength. The classifier with the best performance has been demonstrated to adequately capture the complex nonlinear relationships among hydrochemical parameters and predict drinking water class of groundwater accurately. The comparison establishes a credible standard for estimating groundwater quality using traditional supervised machine learning algorithms. The benchmark results obtained also pave the way for more investigations into explainable artificial intelligence and advanced ensemble learning of groundwater quality assessment.

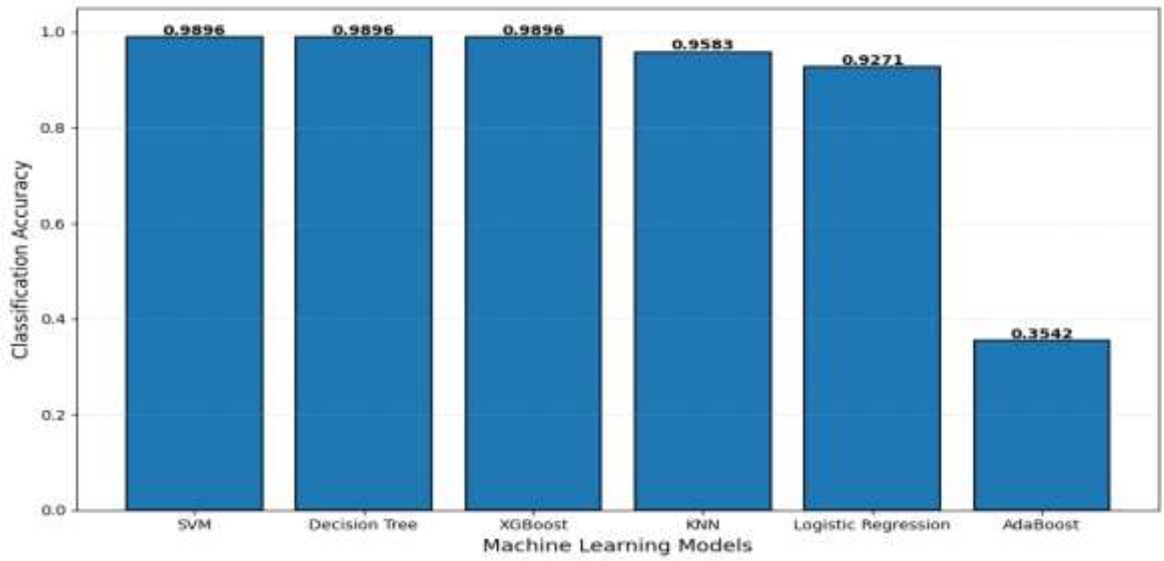

Figure 8. Overall classification accuracy comparison of the evaluated machine learning model

Natural Resources for
Human Health
eISSN: 2583-1194

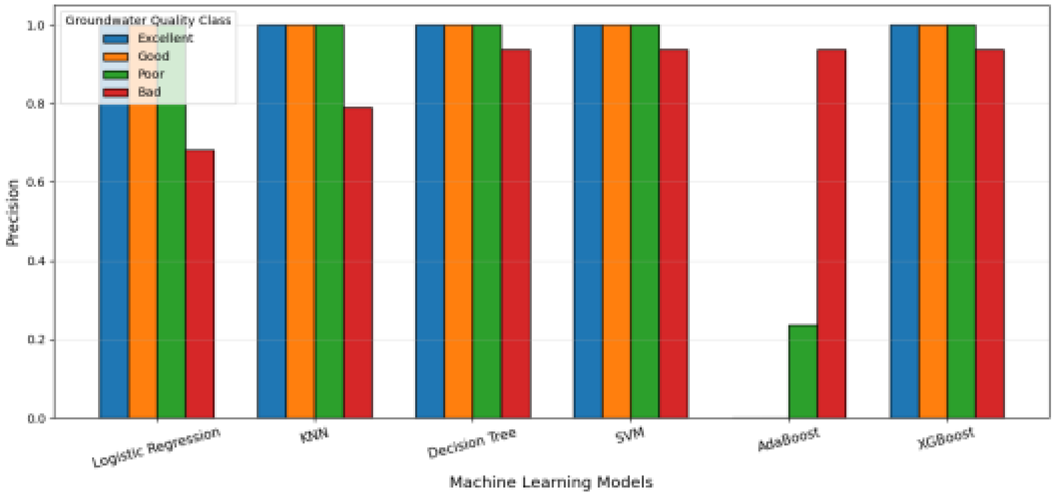

Figure 9. class-wise **Precision** of all the models

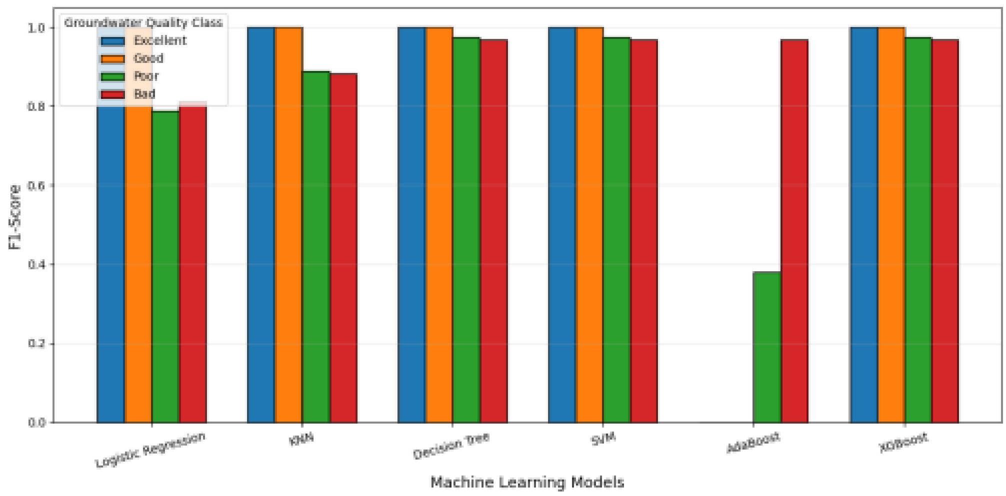

Figure 10. class-wise F1**-Score** of all the models

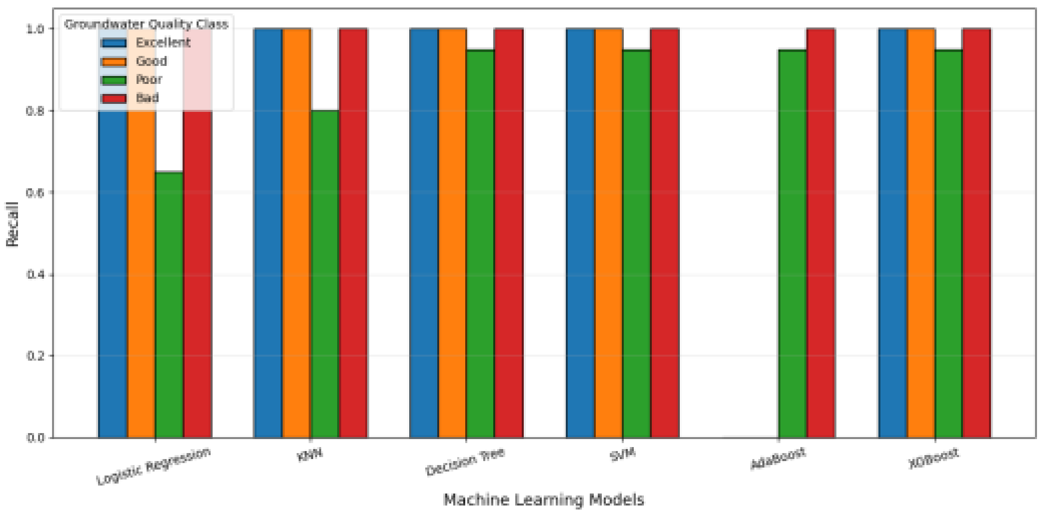

Figure 11. Class-wise Recall Comparison of Machine Learning Models

Natural Resources for Human Health
eISSN: 2583-1194

Table 7. Performance Comparison Metrics of Logistic Regression

Model	Accuracy	Precision	Recall	F1-Score
SVM	98.95	99.02	98.95	98.96
Decision Tree	98.95	99.02	98.95	98.96
XGBoost	98.95	99.02	98.95	98.96
KNN	95.83	96.71	95.83	95.84
Logistic Regression	92.70	95.02	92.70	92.62
AdaBoost	35.41	19.59	35.41	23.03

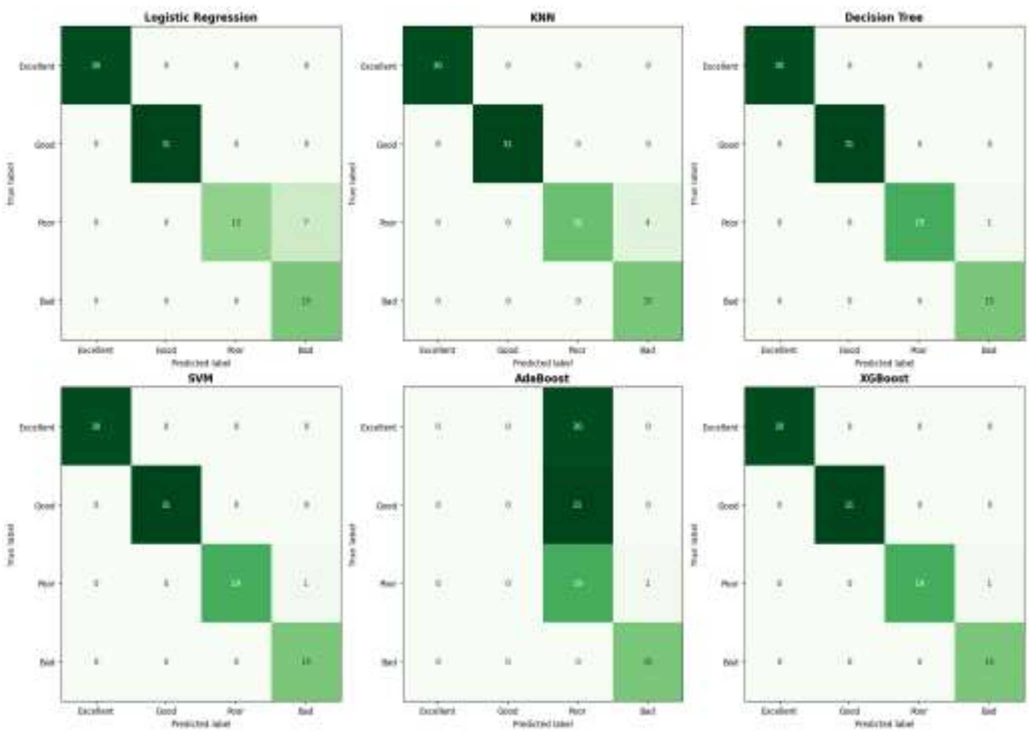

Figure 12. Confusion Matrix for All Models

Further, the confusion matrices of some selected models such as Logistic Regression, KNN, Decision Tree, SVM, Adaboost and XgBoost are presented in Figure 12 to make a holistic observation about the model performance.

5. CONCLUSIONS AND FUTURE SCOPE

5.1 Conclusions

A framework using machine learning for groundwater quality assessment and prediction of its drinkability has been developed based on the study of hydrochemical parameters from the Central Ground Water Board (CGWB), India. The groundwater samples were classified into four drinkability classes, viz Excellent, Good, Poor and Bad, using the Water Quality Index (WQI). Groundwater quality assessment was framed as a multi-class classification problem. Before building the model, exploratory data analysis was carried out to check the statistical characteristics of the dataset, outlier identification, feature correlation analysis, and feature normalization. This was done to enhance quality and reliability of input data. To assess whether conventional supervised learning algorithms can be applied in the study of groundwater quality, six widely used machine learning models were developed. These models are Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support

Vector Machine (SVM), AdaBoost and Extreme Gradient Boosting (XGBoost). Further, they were comparatively examined in terms of performance using similar experimental settings. The models were evaluated for performance using standard multi-class classification metrics (Accuracy, Precision, Recall, F1 measure and Confusion Matrix). The comparison showed that the predictions of ensemble learning methods for groundwater quality assessment are better than the predictions of classical linear, and distance based classifiers. The XGBoost model performed significantly better than all models tested. XGBoost successfully captures complex nonlinear relationships among hydrochemical parameters and classifies groundwater into drinking and non-drinking categories with accuracy. The comparison has shown a reliable benchmark performance for groundwater quality prediction using conventional machine learning methods. Furthermore, data-driven methods can be applied to facilitate groundwater quality monitoring and decision-making. The offered framework is an effective, reliable, and computationally efficient substitute to the laboratory rich conventional groundwater quality assessment procedures. The machine learning model developed can help the environmental agencies and water resource managers for the rapid assessment of groundwater quality and the prioritisation of monitoring for sustainable management of groundwater resources.

5.2 Future Scope

Water quality assessment and prediction for drinking propose have shown a potential framework which can be enhanced in many ways. We aim to gain understanding of the different hybrid ensemble learning frameworks that combined the power of more than one machine learning algorithm for further improvement of prediction accuracy. In particular, advanced boosting and stacking strategies can be investigated for the management of heterogeneous hydrochemical datasets collected from diverse geographical regions. Integrating Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME is another encouraging area that warrants further research for making the groundwater quality prediction models more transparent as well as interpretable. By this, environmental scientists and policy-makers can identify the hydrochemical parameters that contribute most to groundwater quality classification and would help enhance confidence in model predictions. The proposed framework might be enhanced by feeding the spatiotemporal prediction models with geospatial information, remote sensing observations, meteorological variables, land-use characteristics, and temporal groundwater monitoring data to capture rregional and seasonal variations in groundwater quality. By integrating Geographic Information Systems (GIS) with machine learning algorithms will further help to map groundwater quality and assess their risk at larger spatial scales. Further studies would delve into the modeling of complex hydrogeochemical interactions using ANN, LSTM, GNN, and Transformer based approaches using large scale groundwater datasets 7 8 11 12 23 24. When enough training data are available, these sophisticated models may show enhanced prediction performance. Ultimately, the methodology presented in this paper can be integrated within Internet of Things (IoT) based water quality monitoring system and used for real-time assessment of groundwater quality and automatic early warning system. Intelligent monitoring systems can significantly aid in sustainable management of ground water, public health, and evidence-based environmental policy-making. This research establishes a firm baseline for conventional machine learning-based groundwater quality assessment and lays the groundwork for the future development of explainable, intelligent, real-time groundwater quality prediction systems.

REFERENCES

[1] Panigrahi, N., Patro, S. G. K., Kumar, R., Omar, M., et al. (2023). *Groundwater Quality Analysis and Drinkability Prediction using Artificial Intelligence.* Earth Science Informatics, 16, 1701–1725, DOI: 10.1007/s12145-023-00977-z00977-x.

[2] Haggerty, R., Sun, J., Yu, H., & Li, Y. (2023). *Application of machine learning in groundwater quality modeling: A comprehensive review.* Water Research, 233, 119745. https://doi.org/10.1016/j.watres.2023.119745.

[3] Huang, Y., Wang, C., Wang, Y., et al. (2024). *Application of machine learning models in groundwater quality assessment and prediction: Progress and challenges.* Frontiers of Environmental Science & Engineering, 18(3). https://doi.org/10.1007/s11783-024-1789-2.

[4] Torres-Martínez, J. A., Mahlknecht, J., Kumar, M., et al. (2024). *Advancing groundwater quality predictions: Machine learning challenges and solutions.* Science of the Total Environment, 949, 174973. https://doi.org/10.1016/j.scitotenv.2024.174973.

[5] Karunanidhi, D., Raj, M. R. H., Roy, P. D., & Subramani, T. (2025). *Integrated machine learning based groundwater quality prediction through groundwater quality index for drinking purposes in a semi-arid river basin of South India.* Environmental Geochemistry and Health. https://doi.org/10.1007/s10653-025-02425-9.

[6] Abu El-Magd, S. A., Ismael, I. S., El-Sabri, M. A. S., & Abdo, M. S. (2023). *Integrated machine learning–based model and WQI for groundwater quality assessment: ML, geospatial, and hydro-index approaches.* Environmental Science and Pollution Research, 30, 53862–53875. https://doi.org/10.1007/s11356-023-25938-1.

[7] Das, C. R., & Das, S. (2024). *Coastal groundwater quality prediction using objective-weighted WQI and machine learning approach.* Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-024-32415-w.

[8] Groundwater Quality Prediction and Analysis Using Machine Learning Models and Geospatial Technology. Sustainability, 16, 9848 (2024). https://doi.org/10.3390/su16229848.

[9] Hussein, E. E., Derdour, A., Zerouali, B., Almaliki, A., Wong, Y. J., Ballesta-de los Santos, M., Minh Ngoc, P., Hashim, M. A., & Elbeltagi, A. (2024). Groundwater Quality Assessment and Irrigation Water Quality Index Prediction Using Machine Learning Algorithms. *Water*, *16*(2), 264. https://doi.org/10.3390/w16020264.

[10] Md Isa, A. R. ., I. M. . Yusoff, and R. . Abdul Rahman. "Lake Water Quality Assessment Through GIS Based Interpolation Method: A Case Study of Beris Dam, Kedah, Malaysia". *Trends in Sciences*, vol. 21, no. 4, Feb. 2024, p. 7333, doi:10.48048/tis.2024.7333.

[11] KUSHE, V.P., MISHRA, S.S. & CHARHATE, S. Assessment of ground water quality parameters during post monsoon season in three Taluka of Sindhudurg district of Maharashtra using water quality index. *Sādhanā* **49**, 171 (2024). https://doi.org/10.1007/s12046-024-02504-0.

[12] Manna, A., Biswas, D. Assessment of Drinking Water Quality Using Water Quality Index: A Review. *Water Conserv Sci Eng* **8**, 6 (2023). https://doi.org/10.1007/s41101-023-00185-0.

[13] M. Rhishi Hari Raj, D. Karunanidhi, Priyadarsi D. Roy, T. Subramani, Fluoride enrichment in groundwater and its association with other chemical ingredients using GIS in the Arjunanadi River basin, Southern India: Implications from improved water quality index and health risk assessment, Physics and Chemistry of the Earth, Parts A/B/C, Volume 137, 2025, pp. 103765, ISSN 1474-7065, https://doi.org/10.1016/j.pce.2024.103765.

[14] M. Rhishi Hari Raj, D. Karunanidhi, N. Subba Rao, T. Subramani, Machine learning and GIS based groundwater quality prediction for agricultural practices - A case study form Arjunanadi River basin of South India, Computers and Electronics in Agriculture, Volume 229, 2025, pp.109932, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2025.109932.

[15] Meera Rajan, D. Karunanidhi, T. Subramani, B. Preethi, Evaluation of fluoride contamination in groundwater and its non-carcinogenic health hazards in drought-prone river basin of South India, Physics and Chemistry of the Earth, Parts A/B/C, Volume 136, 2024, pp. 103714, ISSN 1474-7065, https://doi.org/10.1016/j.pce.2024.103714.

[16] Mohammad Reza Saadati-Gullojeh, Mojgan Ghanbari, Masoud Salavati-Niasari, Facile preparation and characterization of $Zn_2Ti_3O_8$/g-C_3N_4 nanocomposites for degradation of rhodamine B under simulated sunlight, Solar Energy, Volume 268, 2024, pp. 112316, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2024.112316.

[17] Vinita Sangwan, Rashmi Bhardwaj; Machine learning framework for predicting water quality classification. *Water Practice and Technology* 1 November 2024; 19 (11): 4499–4521. doi: https://doi.org/10.2166/wpt.2024.259.

[18] Jamal Sarafaraz, Fariborz Ahmadzadeh Kaleybar, Javad Mahmoudi Karamjavan, Nader Habibzadeh, Predicting river water quality: An imposing engagement between machine learning and the QUAL2Kw models (case study: Aji-Chai, river, Iran), Results in Engineering, Volume 21, 2024, pp. 101921, ISSN 2590-1230, https://doi.org/10.1016/j.rineng.2024.101921.