Article:

Kushwaha, N.L., Sahoo, M. orcid.org/0000-0003-3552-4691 and Biwalkar, N. (2025) Machine learning approaches for assessing groundwater quality and its implications for water conservation in the sub-tropical capital region of India. Water Conservation Science and Engineering, 10 (1). 25. ISSN 2366-3340

https://doi.org/10.1007/s41101-025-00348-1

Machine Learning Approaches for Assessing Groundwater Quality and Its Implications for Water Conservation in the Sub-Tropical Capital Region of India

Nand Lal Kushwaha [1,2,*], Madhumita Sahoo [3] and Nilesh Biwalkar [1]

[1]Department of Soil and Water Engineering, Punjab Agricultural University, Ludhiana, Punjab, INDIA -141004, Email: nileshbiwalkar-swe@pau.edu (N.B.)

[2]Indian Council of Agricultural Research – Indian Agricultural Research Institute, New Delhi, INDIA-110012,

[3]Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, INDIA-751003, Email: sahoomadhu1989@gmail.com (M.S.)

*Corresponding Author Email: nand.kushwaha@icar.gov.in (N.L.K.)

Abstract

Groundwater is vital for urban areas, serving as a key source of water for domestic, industrial, and agricultural needs. Urban areas face increasing risks of groundwater contamination due to growing reliance on groundwater, with pollution arising from intensified human activity, including sewage leaks, industrial waste, and improper waste disposal. Consequently, assessing groundwater quality has become essential for ensuring sustainable water management. The present study aims to develop and evaluate four machine learning models, namely Support Vector Machine (SVM), Random Forest Model (RFM), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGB), for groundwater quality prediction and develop spatial groundwater quality maps to guide conservation efforts for the highly polluted and urbanised National Capital Territory (NCT), Delhi, India. The model performances were assessed using six statistical indicators i.e., Willmott's Index (WI), Nash Sutcliffe model Efficiency coefficient (NSE), Percent bias (PBIAS), Mean absolute error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) and graphical representation i.e., radar chart and Taylor diagram. Results revealed that the performance of the RFM model (WI = 0.850, NSE = 0.947, R^2 = 0.938, PBIAS = 12.024, MAE = 45.912, and RMSE = 111.436) was superior to the SVM, GBM and XGB models for prediction of GWQI. Interestingly, the SVM model

shows significantly worse performances in predicting the GWQI. The outcomes of the present study will provide valuable insights for water policymakers, offering groundwater quality information to guide sustainable groundwater management and conservation efforts.

Keywords: Water conservation; GWQI; Random Forest; Urban water management; Taylor diagram.

Abbreviations	
NCT	National Capital Territory
SVM	Gradient Boosting Machine
GBM	Reduced Error Pruning Tree
RFM	Random Forest Model
XGB	Extreme Gradient Boosting
EC	Electrical conductivity
Ca^{2+}	Calcium
Mg^{2+}	Magnesium
Na^{+}	Sodium
K^{+}	Potassium
Cl^{-}	Chloride
CO_3^{2-}	Carbonate
HCO_3^{-}	Bicarbonate
SO_4^{2-}	Sulphate
NO_3-	Nitrate
TH	Total Hardness
GWQI	Ground Water Quality Index
WI	Willmott Index
R^2	Coefficient of Determination
PBIAS	Percent Bias
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
NSE	Nash–Sutcliffe Efficiency
GWYB	Groundwater Year Book
CGWB	Central Ground Water Board
SDGs	Sustainable Development Goals
ANN	Artificial Neural Networks
M5P	M5Preuning Tree
MARS	Multivariate Adaptive Regression Splines
ELM	Extreme Learning Machine
GEP	Gene Expression Programming
MSL	Mean Sea Level
WHO	World Health Organization
µS/cm	Micro-siemens per Centimeter
mg/l	Milligram per Liter

1. Introduction

Groundwater is a ubiquitous and reliable source of potable water. Reliance on groundwater is due to its superior quality in comparison to surface water [1]. Groundwater plays a vital role in sustainable urban development worldwide. With surface water sources increasingly contaminated by human activities, urban areas now rely heavily on groundwater, supplying over half the potable water in many Asian cities [2, 3]. However, groundwater quality vulnerability has increased due to undesirable recharge from underground storage reservoirs (of volatile organic compounds) and sewer systems, overexploitation of groundwater, and surface-subsurface interaction with contaminated urban streams. Urban aquifers in industrialized countries often suffer from contamination due to exposure to petroleum hydrocarbons, chlorinated solvents, pesticides, and improper waste disposal [4]. The global urban population may add another 2.5 billion people to the existing urban population by 2050 [5]Addressing groundwater quality in an urban context is, therefore, of paramount importance for the efficient management of the subsurface water resource.

Vulnerability assessments serve to direct groundwater protection efforts in a way that the most environmental and public health benefits are achieved at least cost [6–8]. Groundwater quality index (GWQI) is an efficient method of classification of water for their suitability/unsuitability for human consumption. GWQI is a dimensionless index calculated from selected water quality parameters [3, 9]. Water quality parameters are assigned weights depending on their impact on water quality and a combined quality index serves as a GWQI. Since different water quality parameters can vary differently and have an unpredictable impact on the overall quality of groundwater, it is necessary to use algorithms/methods that can ease the calculation of GWQIs. A machine learning algorithm learns from patterns in input data and adjusts to improve estimated output [10, 11]. The broad range of algorithms under the umbrella of machine learning have found applications in most scientific disciplines [12, 13]. Processes that exhibit a high-level of non-linearity and uncertainty, can make use of machine learning algorithms for prediction and analyses. Groundwater quality modeling is an excellent example of one such process, where a high degree of temporal and spatial variability exists. Machine learning algorithms have been known to give excellent results in highly non-linear systems. These algorithms can handle multidimensional and multivariate calculations efficiently. Their applications have been on significant increase due to their ease of handling and the various other advantages over traditional statistical methods [11, 14]. Machine learning algorithms have been used to predict groundwater contamination levels [15, 16], mapping groundwater quality [17], and predicting groundwater quality status [18]. Researchers have been using GWQI with combinations of different methods to assess an overall picture of groundwater quality in their study areas. A brief overview on various methods used for developing GWQI in the past few decades has been summarized in Table 1. In this study, four machine learning models namely Support Vector Regression (SVM), Random Forest Model (RFM), Gradient Boosting Mechanism (GBM), and Extreme Gradient Boosting (XGB) were developed and evaluated to determine the best model that can provide a dependable GWQI for the current urban setting.

Recent studies have demonstrated the potential of machine learning in predicting water quality index (WQI) with high accuracy and efficiency. Mamat et al. [19] and Tabassum et al. [20] both utilized SVM and other machine learning techniques, respectively, to enhance WQI prediction. Mamat et al. [19] explored the exceptional ability to replicate the Department of Environment (DOE)-WQI and Tabassum et al. [20] addressed the limitations of traditional approaches through machine learning-based WQI prediction models. Goodarzi et al. [21] further explored the use of three machine learning models in estimating WQI, with the Multivariate Adaptive Regression Splines (MARS) model being slightly more accurate. Yadav et al. [22] extended the application of machine learning to predicting influent and effluent quality parameters in a wastewater treatment plant, achieving a strong correlation between measured and predicted parameters. Khoi et al. [23] found that XGB outperformed other models in predicting WQI in the La Buong River, Vietnam. Similarly, Bui et al. [24] further improved WQI prediction using hybrid machine learning models, with the Bagging (BA) and found that the BA-RT (Random Tree) model performed the best. Ganga Devi [25] and Sakaa et al. [26] also found that RFM has the potential ability to predict the water quality index (WQI). Nayan et al. [27] showed excellent agreement between predicted and observed water quality by GBM. Osman et al. [28] compared the performance of XGB, ANN and SVM and found that the XGB model outperformed both the SVM and ANN models. Similarly, Mo et al. [29] applied XGB and RFM for WQI prediction. They concluded that both models provided the most accurate WQI predictions, especially in winter, using minimal key parameters like Ammonia Nitrogen, Total Phosphorus, Dissolved Oxygen, and turbidity. Accuracy exceeded 80% for good predictions in spring and winter but dropped to 70% in summer and autumn. Mohseni et al [2] conducted study to predict the Urban water quality index (WQI) for Ujjain city, Madhya Pradesh, India, using four machine learning models (ANN, SVM, RF, and XGB) along with multiple linear regression (MLR). Among the models, XG-Boost outperformed others, achieving the highest accuracy with $R^2 = 0.987$, RMSE = 3.273, and MAE = 2.727 during testing, and an AUC of 0.9048 validated its robustness. These studies collectively highlight the potential of machine learning, particularly RFM, XGB, GBM, and hybrid models for reliable WQI predictions, aiding decision-makers in urban water management. Reliable, generalizable, and stable models are needed to anticipate water quality parameters in real-time. Even when they perform well generally, certain models may not be appropriate for prediction, because of their great sensitivity to the input variables. Therefore, the stability of the machine learning (ML) models in the forecasting of the water quality parameters in real time is critical.

Table 1 Previous methods applied in developing GWQI

Studies	Year of study	Region	Method applied
Saeedi et al. [30]	2010	Qazvin plateau area, Iran	Principal Component Analysis
Vasanthavigar et al. [31]	2010	Thirumanimuttar sub-basin, India	Laboratory analysis
Rajankar et al. [32]	2011	Bhandara district, India	Statistical analysis
Sadat-Noori et al. [33]	2014	Saveh-Nobaran plain, Iran	Geographical Information System (GIS)
Batabyal & Chakraborty, [34]	2015	Bardhaman District, India	Laboratory analysis and GIS
Dhar et al. [35]	2015	Kanpur, India	Multi-criteria decision analysis and GIS
Varol & Davraz, [36]	2015	Tefenni (Burdur) plain, Turkey	Laboratory analysis and statistical analysis
Boateng et al. [37]	2016	Ejisu-Juaben Municipality, Ghana	Laboratory analysis and statistical analysis
Adimalla & Taloor, [1]	2020	Medak, India	Piper Trilinear diagram and Gibbs diagram
Norouzi & Moghaddam, [38]	2020	Miandoab plain aquifer, Iran	Machine learning models
Fang et al. [39]	2020	Dagu river basin, China	Statistical analysis
Ram et al. [40]	2021	Mahoba district, India	Hill-Piper Trilinear diagram
Singha et al. [41]	2021	Mahanadi basin, India	Machine learning models
Raheja et al. [42]	2021	Haryana, India	Machine learning models
Mozaffari et al. [43]	2022	Zanjan province, Iran	Machine learning models
Dimple et al. [44]	2022	Nand Samand catchment, India	Data-driven models
Kushwaha et al. [16]	2023	Pusa Campus, New Delhi, India	Machine learning models
Mamat et al. [19]	2023	Langat River catchment, Malasiya	Data-driven modeling
Goodarzi et al. [21]	2023	Yazd-Ardakan Plain, Iran	Machine learning models
Mohseni et al. [2]	2024	Ujjain, Madhya Pradesh, India	Machine learning models
Saha et al. [45]	2024	Ganges delta, Indo-Bangladesh region	Machine learning models

As evident from Table 1, the application of machine learning models in developing Groundwater Quality Indices (GWQI) has gained significant traction over the past decade. These models have demonstrated considerable potential in tackling diverse challenges associated with groundwater quality prediction and assessment. For instance, Support Vector Machines (SVM) have been effectively employed for groundwater quality mapping [46], groundwater quality prediction [47], and spatial analysis of a groundwater quality parameter [48]. Similarly, the RFM model has successfully predicted groundwater quality [49], and groundwater vulnerability assessment [50]. Gradient Boosting Machines (GBM) and XGB models have been used in conjunction with other machine learning models in groundwater quality prediction [51–53].

Despite these advancements, research gap persists, particularly regarding machine learning-based water quality prediction in the NCT, Delhi. The absence of comprehensive studies underscores the urgent need for in-depth investigations to tackle the critical water quality challenges in one of the most densely populated and rapidly urbanizing regions in India. The novelty of this study lies in its tailored application of ML models to assess groundwater quality in the sub-tropical capital region of India, a critical area lacking systematic evaluations despite its vulnerability to overexploitation and pollution. The present study aims to address these gaps by developing and evaluating robust machine learning models for the prediction of groundwater quality (GWQI) for NCT, Delhi, a region facing significant groundwater quality challenges. The present study systematically developed, evaluates and compares the performance of multiple machine learning models (SVM, RF, GBM, XGB) to identify the most effective approach for GWQI prediction offering insights into their practical implications for groundwater quality assessment and water conservation planning.

2. Material and Methods

2.1. Study Area and Available Datasets

The National Capital Territory (NCT) of Delhi covers 1483 km^2. It is located between the latitude of 28° 24' 15" and 28° 53' 00" N and longitude of 76° 50' 24" and 77° 20' 30" E (Fig. 1). The observations for the present study were obtained from Central Groundwater Board (CGWB) state unit office, Delhi through Groundwater Year Book (GWYB) for 2020. According to the 2011 census, the population of NCT Delhi is 167.87 lakhs, with a population density of 11320 per km^2. The average annual rainfall of the NCT of Delhi is 611.8 mm. The monsoon season from July, through September receives about 80% of the yearly rainfall. Long-term rainfall data from 1984 to 2017 reveal that Delhi's rainfall is very varied, which in turn influence the ground water's natural replenishment each year [54] . The diverse geological formations of NCT Delhi have hydrogeological characteristics that Delhi quartzite and older and younger alluvium, regulate the availability of groundwater. The CGWB has installed monitoring stations located throughout both the alluvial and quarzitic areas of the NCT of Delhi and monitoring the groundwater level and quality at regular intervals.

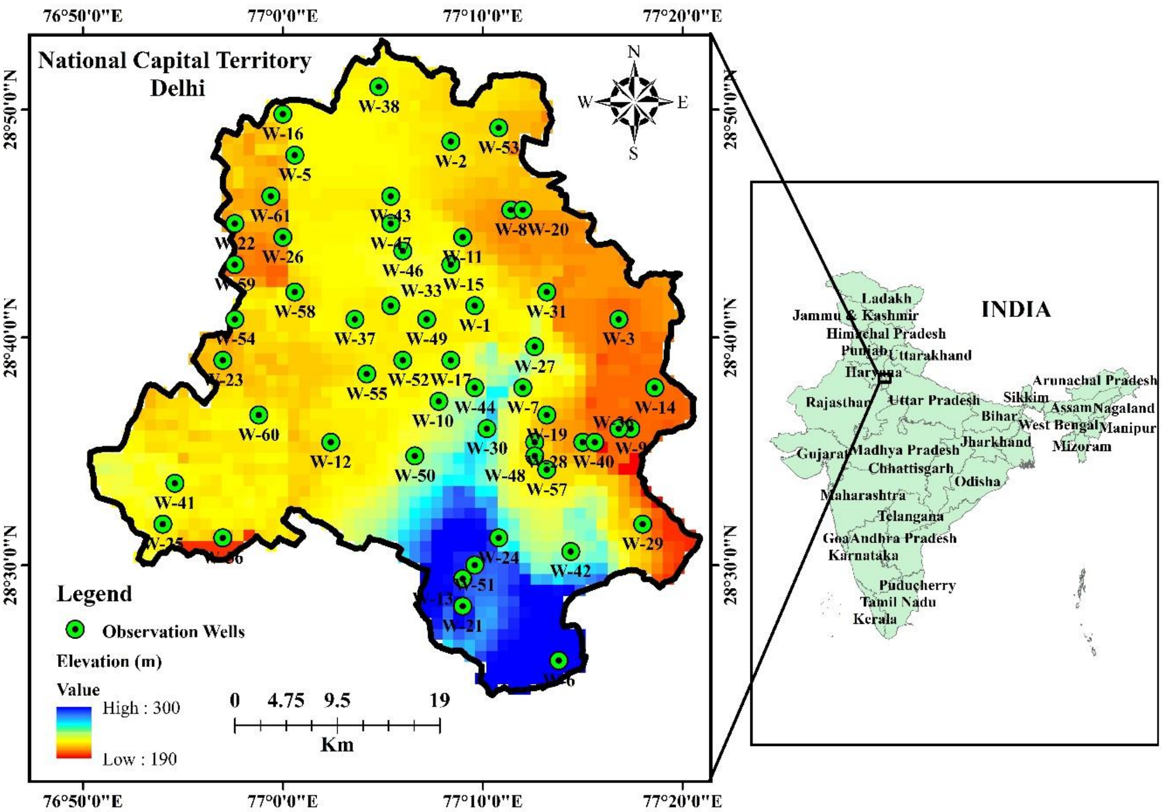

Fig. 1 Location map of the study area showing the observation well location within the NCT Delhi.

Ground water quality datasets on electrical electrical conductivity (EC), carbonate (CO_3^{2-}), bicarbonate (HCO_3^-), chloride (Cl^-), Sulphate(SO_4^{2-}), nitrate (NO_3^-), fluoride (F^-), phosphate (PO_4^{3-}), calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^+), potassium (K^+), silicon dioxide (SiO_2), and total hardness (TH) were obtained from the Ground Water Year Book, CGWB, National Capital Territory, Delhi (http://cgwb.gov.in). Most of the eastern part of NCT Delhi, in areas around the Yamuna flood plain and Delhi Quartzite Ridge zones, has EC within the permissible range of 0 to 2250 µS/cm at 25°C whereas rest of NCT Delhi, except some pockets of South West, North West and West District, has EC value of more than 3000 µS/cm at 25 °C. It is also observed that water from deeper aquifers has greater EC value than the water from shallow aquifers. The detailed methodology for the present study is presented in the Fig. 2.

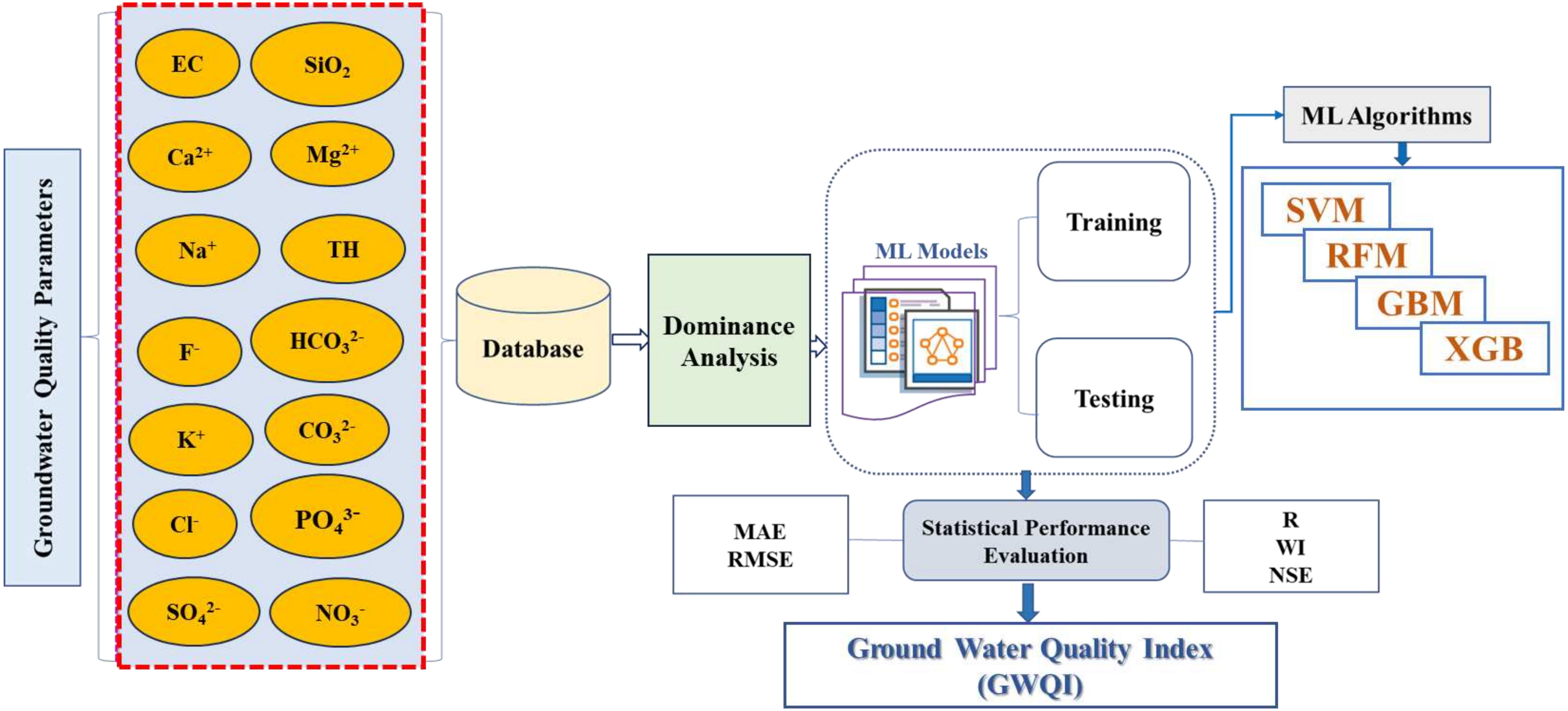

Fig. 2 Ground water quality prediction using machine learning models

2.2. Data Analysis

2.2.1 Computation of the groundwater quality index (GWQI)

The quality of the groundwater for drinking purposes was determined based on the values of the GWQI. The GWQI was computed by assigning specific weight to individual physicochemical parameters [55]. The GWQI is defined as a rating that reflects the composite influence of different physicochemical parameters of water [56, 57]. It is an important tool for determining water quality for drinking purposes. GWQI was calculated using the following steps -

1. Each one of the 14 water quality parameters was assigned "weight" number (W_i). These numbers describe the significance of parameters in classifying the suitability of water for drinking purposes. Mineralization, SO_4, Cl, and F are assigned the highest rating of "5" due to their direct impact on water quality and human health [57, 58]. The CO_3 and HCO_3, on the other hand, have assigned a minimum value of "1".
2. "Relative weight" (W_r) of each physicochemical parameter was determined using equation (1). The assigned weights (W_i), relative weights (W_r), and the WHO standard have been given in Table 2.

$$W_r = \frac{W_i}{\sum_{i=1}^{n} W_i} \tag{1}$$

where W_r is the relative weight of the i^{th} parameter; W_i is the weight assigned to i^{th} parameter and n is the number of parameters.

3. "Quality rating" (q_i) for each parameter was determined using the equation (2)

$$q_i = \frac{c_i}{s_i} \tag{2}$$

where, q_i is the quality rating, c_i is the chemical concentration (mg/l), and s_i is the WHO drinking water quality standard (mg/l) of i^{th} parameter.

4. Calculate GWQI using equation (3)

$$GWQI = \sum_{i=1}^{n} W_r \times q_i \tag{3}$$

2.2.2. Data preprocessing

The different water quality parameters have different ranges. All the data were, therefore, scaled between [0,1] using Equation (1). The x_{min} and x_{max} are minimum and maximum values, respectively, of the specific parameter in the data that is being scaled. The scaled parameters and outputs from the analysis are rescaled back to the original values after the analysis were completed.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4}$$

The data were subjected to preliminary analyses using Boxplot graphs (Fig. 3) and Pearson correlation matrix analysis (Fig. 4). The results of these preliminary analyses helped in identifying the correlated parameters. Further, the parameters were subjected to dominance analysis.

Table 2 Details of physical and chemical water quality parameters, assigned weights and relative weights based on the WHO standard

S. No.	**Water quality parameter** (All parameter measured in mg/l, except for EC)	**Weight (W_i)**	**Relative weight (W_r)**	**WHO Standard** [59]
1	EC (µS/cm at 25^0 C)	2	0.046512	400
2	CO_3^{2-}	1	0.023256	80
3	$HC0_3^-$	1	0.023256	300
4	Cl^-	5	0.116279	250
5	SO_4^{2-}	5	0.116279	200
6	NO_3^-	4	0.093023	42
7	F^-	5	0.116279	1
8	PO_4^{3-}	3	0.069767	30
9	Ca^{2+}	3	0.069767	75
10	Mg^{2+}	3	0.069767	30
11	Na^+	4	0.093023	200
12	K^+	2	0.046512	20
13	SiO_2	2	0.046512	20
14	TH	3	0.069767	300
		$\sum W_i = 43$	$\sum W_r = 1$	

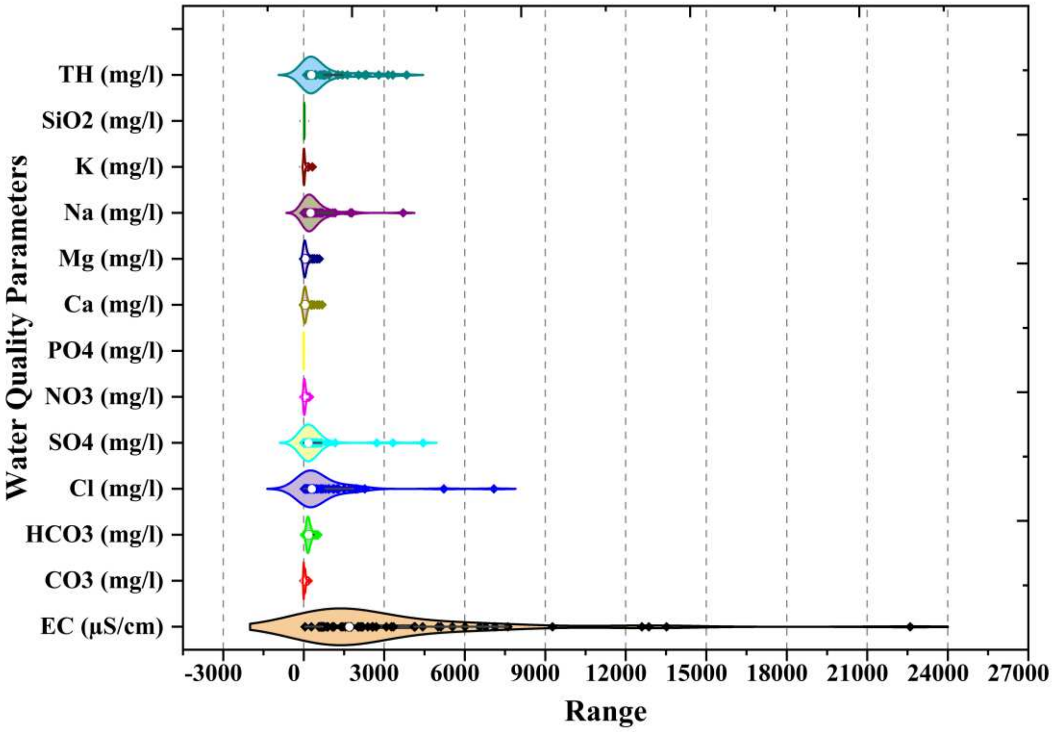

Fig. 3 Violin plot showing the distribution of water quality dataset used for model development

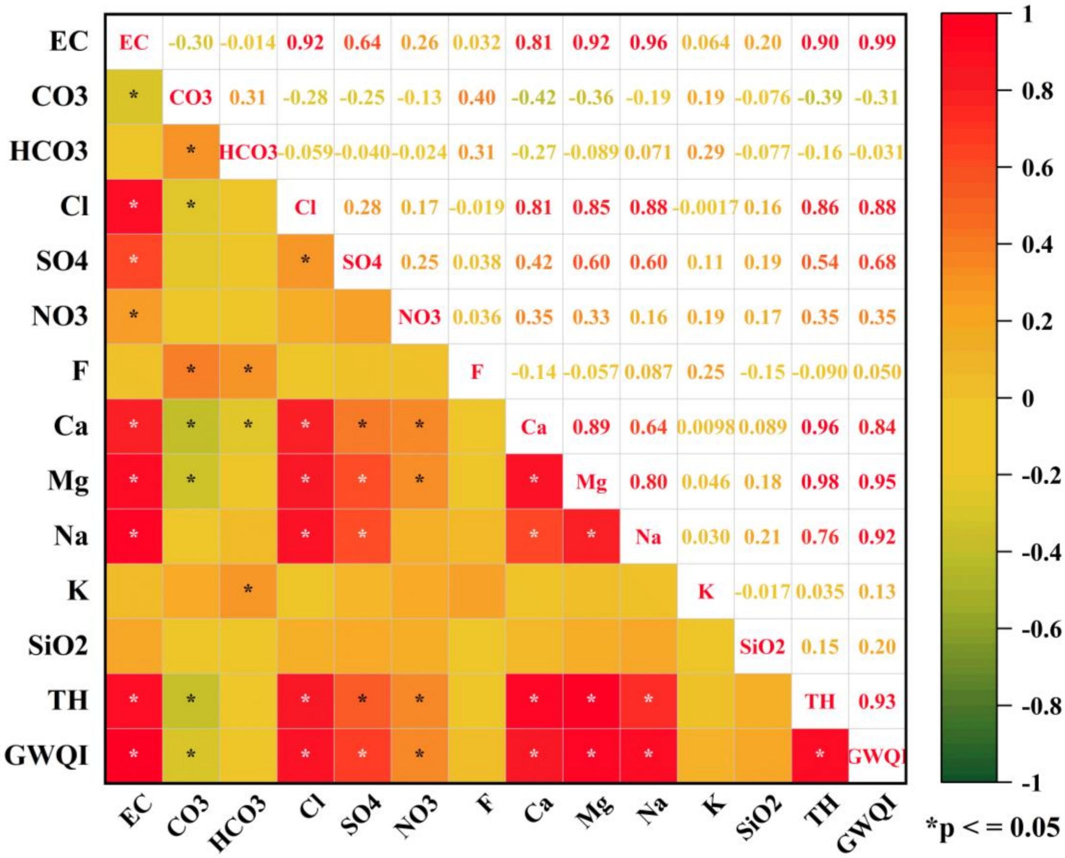

Fig. 4 Pearson correlation matrix analysis. Warm and cold colours indicate positive and negative correlations, respectively, and darker colours indicate stronger correlations.

2.2.3. Dominance analysis

The calculation for GWQI can simplified by taking up the relatively more influential water quality parameters than the entire set of parameters. The relative importance of parameters can be determined by performing dominance analysis [60]. Dominance analysis determines the relative importance of one independent variable over other independent variables in multiple regression. Based on the coefficient of determination, R^2, between the dependent and independent variables, the ranking of individual variables is obtained. The selection of the most influential parameters can be done from this rank list. In the present study, EC, Cl^-, SO_4^{2-}, NO_3^-, Ca^{2+}, Mg^{2+}, Na^+, and TH were found to be relatively most influential parameters for the calculation of GWQI. Dominance analysis was performed within R environment (R Core Team, 2022) using domir (Luchman, 2022) package.

The dataset was separated into training (70%) and testing data (30%), before analysing for regression using multiple machine learning models.

2.3. Machine Learning Algorithms

2.3.1 Support vector machine

Support vector machine (SVM) is a robust algorithm based on the structural risk minimization principle to handle complex non-linear problems with ease [61]. The SVM algorithm aims at finding the best fit hyperplane within an n-dimensional space to predict values with a minimum error. A hyperplane is a decision boundary line with the maximum number of data points. Non-linearity is handled by using kernel functions in SVM. Kernel functions transform the input data into a desired form to search for a hyperplane. A discussion on SVM models as derived from Chervonenkis [62] is given as follows:

For a dataset as $\{(\alpha_1, \beta_1), \ldots\ldots\ldots, (\alpha_n, \beta_n)\} \subset \aleph \times \Re$ where $\aleph$ denotes the space of the input patterns, α_i and β_i are the predictor and response variables, respectively. In the SVM, the goal is to find a function $f(\alpha)$ that has the most ε deviation from the actually obtained targets β_i for all the training data and at the same time as flat as possible [16, 63–65]. Smola & Schölkopf (2004) described the basic equation taking the case of linear f by taking the form

$$f(\alpha) = \langle \omega, \alpha \rangle + b \tag{5}$$

with $\omega \in \aleph$, $b \in \Re$. Eq. (2) gives the prediction for a given sample. A minimization of ω will suffice for flatness. The convex optimization problem for minimizing ω is as follows-

$$\min_{\omega \in \Re} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right) \tag{6}$$

subject to

$$\begin{cases} \beta_i - \langle \omega, \alpha_i \rangle - b \le \varepsilon + \xi_i \\ \langle \omega, \alpha_i \rangle + b - \beta_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

The cost constant C, the ε amount of deviation, and the kernel parameters control the parameters of SVM [63–65]. The SVM offer high accuracy in classification tasks by identifying optimal hyperplanes, especially effective in high-dimensional spaces. SVMs excel in handling non-linear relationships through kernel functions, providing flexibility in capturing complex patterns. However, SVMs may struggle with large datasets due to their computational intensity. They are also sensitive to the choice of kernel and parameter settings, requiring careful tuning for optimal performance. In the present study, radial basis function (RBF) kernel was used. The SVM analysis was performed in R environment using e1071 [67] and other supporting packages. The values and ranges of hyperparameters are presented in Table 3.

Table 3 Model architecture used in the machine learning models

Machine learning models	**Hyperparameter range/value/name**
Support Vector Machine	

ε	0.1
C	2^0 - 2^{10}
Kernel function	RBF
Random Forest Model	
n_{tree}	50
m_{try}	5564
node size	4
Gradient Boosting Mechanism	
n.trees	1716
n.minobsinnode	2
distribution	Gaussian
shrinkage	0.67
EXtreme Gradient Boosting	
nrounds	247
maximum depth	5
η	0.353
λ	0.78

2.3.2 Random forest model

Random forest model (RFM) is a supervised machine learning method that depends on an ensemble of predictions made by multiple subsets of the given dataset [68] . The training dataset is randomly split into multiple training subsets (individual decision trees). n_{tree} denotes the number of trees to grow in the forest [63, 64, 68]. Each decision tree will generate an output based on an independent training. The response data for each tree are split into two descendant nodes to maximize homogeneity. Random sample of predictors (m_{try}) are chosen and the best split is selected among these variables [64]. A final output is predicted by averaging all the predictions obtained from each sample. Each descendant node of the selected split is treated similarly as the original node and the process is continued repeatedly until a stopping criterion is met. 'Node size' parameter determines minimum size of terminal nodes of the decision trees. RFM offer robustness and improved accuracy by aggregating predictions from multiple decision trees, reducing the risk of overfitting. Additionally, they excel in handling large and complex datasets, providing feature importance rankings that aid in insightful data analysis. In the present study, RFM analysis was performed using randomForest [69] package in the R environment. Model parameters are presented in Table 3.

2.3.3 Gradient boosting machine

Gradient boosting machine (GBM) is another ensemble method like RFM, which creates multiple weak or "poor" performing models and combines them with "strong" models to obtain highly accurate prediction [64, 70]. The initial model will give a "poor" prediction with high prediction errors. These prediction errors from each step are to be minimized to obtain a better prediction. Prediction errors from each step are scaled between [0,1] and then added/combined to the previous prediction to reduce the error. At each step, a new prediction tree is created. Iterations in these steps continue until improvement in prediction is stopped. *n.trees* denotes the total number of trees to fit. *n.minobsinnode* specifies the minimum number of observations in the tree terminal nodes. The *shrinkage* parameter decides the learning rate in the algorithm. The predictor and response datasets are related to each other with some probabilistic distribution. In the present study, Gaussian distribution was assumed between the datasets (Table 3). In the R environment, gbm (Ridgeway et al., 2015) package was used to perform this regression analysis.

2.3.4 EXtreme gradient boosting

EXtreme gradient boosting (XGB) is a supervised learning algorithm similar to GBM in using regression trees as their base estimators. However, the approaches to create trees and to determine splits are different. Another difference between XGB and GBM lies in the inclusion of a regularization hyperparameter (λ) in the output. In XGB's architecture, *nrounds* shows the maximum number of iterations. The maximum depth hyperparameter controls the depth of the tree. The greater the maximum depth the less stable the model becomes. η is the learning rate of the model (Table 3). XGB gives a faster solution convergence than GBM. In R, the xgboost (Chen et al., 2015) package was used to perform XGB regression analysis.

2.4. Statistical Performance Indicators

The performance of the models was evaluated qualitatively through visual observation and quantitatively through the application of various statistical criteria, including the Willmott Index (WI), Nash Sutcliffe model Efficiency coefficient (NSE), Percent bias (PBIAS), Mean absolute error (MAE), Root Mean Square Error (RMSE) and coefficient of determination (R^2). These statistical parameters are summarized in Table 4. In addition to the statistical parameters, the correctness of the investigated models was validated using Box-and-whisker plots, radar chart and a Taylor diagram (TD) [71], among other techniques (i.e., scatter plot). A simplified definition of the Taylor diagram is a thorough depiction of the observed and expected data [72, 73]. Taylor delivered a single demonstration that demonstrated how to show several assessment metrics in real time, at the same time. Correlation coefficients and standard deviation values between predicted and observed values might be shown in this diagram to aid in the detection of changes between the two values [74]. All parameters are specified as follows: $GWQI_A^i$ is the recorded or actual value; $GWQI_P^i$ is the estimated or predicted value, $\overline{GWQI_A^i}$ and $\overline{GWQI_P^i}$ are the mean values of recorded and estimated samples, and N is the total number of selected samples (Table 4)

Table 4 Statistical performance indicator used for model correctness

Equation	**Range**	**Ideal value**	**References**
$\mathrm{PBIAS} = \frac{\sum_{i=1}^{n}\left(\mathrm{GWQI_A^i} - \mathrm{GWQI_P^i}\right)}{\sum_{i=1}^{n} GWQI_A^i} \times 100]$	-∞ to ∞	0	[75, 76]
$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{GWQI_A^i} - \mathrm{GWQI_P^i})^2}$	0 to ∞	0	[77, 78]
$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert\mathrm{GWQI_P^i} - \mathrm{GWQI_A^i}\rvert$	0 to ∞	0	[79, 80]
$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{N}\mathrm{GWQI_A^i} - \mathrm{GWQI_P^i})^2}{\sum_{i=1}^{N}(\mathrm{GWQI_A^i} - \overline{\mathrm{GWQI_A^i}})^2}\right]$	-∞ to 1	1	[81, 82]
$\mathrm{WI} = 1 - \frac{\sum_{i=1}^{N}(\mathrm{GWQI_A^i} - \mathrm{GWQI_P^i})^2}{\sum_{i=1}^{N}\left(\lvert\mathrm{GWQI_P^i} - \overline{\mathrm{GWQI_A^i}}\rvert + \lvert\mathrm{GWQI_A^i} - \overline{\mathrm{GWQI_A^i}}\rvert\right)^2}$	0 to 1	1	[82]
$\mathrm{R^2} = 1 - \frac{\sum_{i=1}^{N}(\mathrm{GWQI_A^i} - \mathrm{GWQI_P^i})^2}{\sum_{i=1}^{N}(\mathrm{GWQI_A^i} - \overline{\mathrm{GWQI_P^i}})^2}$	0 to 1	1	[83, 84]

3. Results and Discussion

3.1. Data Analysis using Inter Correlation Matrix

The descriptive statistical characteristics of the water quality parameters are shown in Table 5. The correlation among the water quality parameters for all observation wells and the importance evaluation of input variables have been carried out using SPSS software (version 17.0) (Fig. 4). The correlation is said to be strong if the correlation coefficient (r) is greater than 0.9, and good if the r varies between 0.75 and 0.9. Similarly, if the value of r is between 0.6 and 0.75, the correlation is said to be moderate and if the value of r is less than 0.6, the correlation is regarded as weak [14, 57, 85]. The analysis shows that the electrical conductivity (EC) has strong correlations with TH (0.9), Na (0.96), Mg (0.92) and Cl (0.92) parameters; and it has good correlation with Ca (0.81). Similarly, the Cl has good correlations with Ca (0.81), Mg (0.85), Na (0.88) and TH (0.86). The F, NO_3, K, SiO_2, CO_3 and HCO_3 have weak correlation with all parameters. It also found that CO_3 has the lowest correlation with all other water quality parameters. The GWQI, which is the target class of the present study has a strong co-relation with EC (0.99), Cl (0.88), Mg (0.95), Na (0.92), and TH (0.93). These findings indicate that F, NO_3, K, SiO_2, CO_3 and HCO_3 have no significant correlation with GWQI.

Table 5. Descriptive statistical characteristics of the used water quality parameter

Water quality parameter	**Mean**	**Standard Deviation**	**Skewness**	**Kurtosis**	**Minimum**	**Q1**	**Median**	**Q3**	**Maximum**
EC (µS/cm)	3200.36	4010.27	2.73	9.16	50	699	1700	4120	22600
CO_3 (mg/l)	30.08	35.65	1.33	2.30	0	0	37	50	164
HCO_3 (mg/l)	196.18	99.58	1.33	2.05	25	139	176	239	528
Cl (mg/l)	714.18	1182.07	3.72	16.48	20	104	299	813	7087
SO_4 (mg/l)	381.57	767.33	4.01	17.05	0	41	171	366	4451
NO_3 (mg/l)	56.66	62.66	1.31	0.62	0.34	7.15	38	81	217
F (mg/l)	0.87	0.68	1.11	0.96	0	0.325	0.6	1.25	3.1
PO_4 (mg/l)	0.10	0.00	1.03	-2.07	0.1	0.1	0.1	0.1	0.1
Ca (mg/l)	103.87	144.97	2.42	5.61	8	24	45	86	681
Mg (mg/l)	108.70	135.37	1.90	2.99	12	25	42	132	579
Na (mg/l)	439.69	597.06	3.38	14.82	9	103	255	498	3703
K (mg/l)	26.12	52.48	3.96	17.81	1	3	9	22	320
SiO_2 (mg/l)	21.48	3.25	-0.21	-0.03	14	19	22	23	29
TH (mg/l)	703.56	895.34	2.01	3.26	102	204	286	756	3833

3.2. Prediction of Groundwater Quality

3.2.1. Training of applied ML Models

To train the applied models, the selected raw data were normalized, scaled 0 to 1 and separated into two datasets: 46 samples were used to train the models and 15 samples were used for testing the models. In the present study, EC, Cl^-, SO_4^{2-}, NO_3^-, Ca^{2+}, Mg^{2+}, Na^+, and TH were found to be relatively the most influential parameters for the prediction of GWQI. The models were developed as per the model architecture in Table 3 in the R environment. The obtained results from the training of models are presented in Table 6. This revealed that GBM and XGB are comparable in the prediction of GWQI. However, the GBM model performed than XGB with high values of WI (0.999), NSE (0.998) and R^2 (0.998); lower values of PBIAS (0.003), MAE (0.079) and RMSE (0.097) in the prediction of GWQI. It was followed directly by the XGB model which has R^2 = 0.992, WI = 0.999, NSE = 0.995, MAE = 2.029, and RMSE = 3.260. The SVM was one of the lowest-performing model in the training phase and it has R^2 = 0.936, WI = 0.945, NSE = 0.830, PBIAS = 0.000, MAE = 68.103, and RMSE = 75.884. The analysis of performance indicators showed that all four models performed well during the training phase. The comparison of observed and predicted GWQI for the selected models were compared graphically presented as scatter plots (Fig. 5). The accuracy of the models is satisfactory when the values are distributed over or evenly on both slides the 1:1 line, showing that the errors obey the Gaussian distribution. From Fig. 5, the predicted values from the XGB (R^2 = 0.992) and GBM (R^2 = 0.998) models are more closely distributed along the 1:1 line compared to the RFM and SVM models.

Furthermore, the comparison among the predicted GWQIs was carried out through a line plot (Fig. 6) and radar chart (Fig. 7 a) between the computed and predictive value of GWQI. This also reflects that the GBM and XGM lines overlapped on the computed GWQI. This confirms the efficacy of both models in the prediction of GWQI. The GBM and XGB model demonstrated the best statistical measures, reflecting their superiority over the other models.

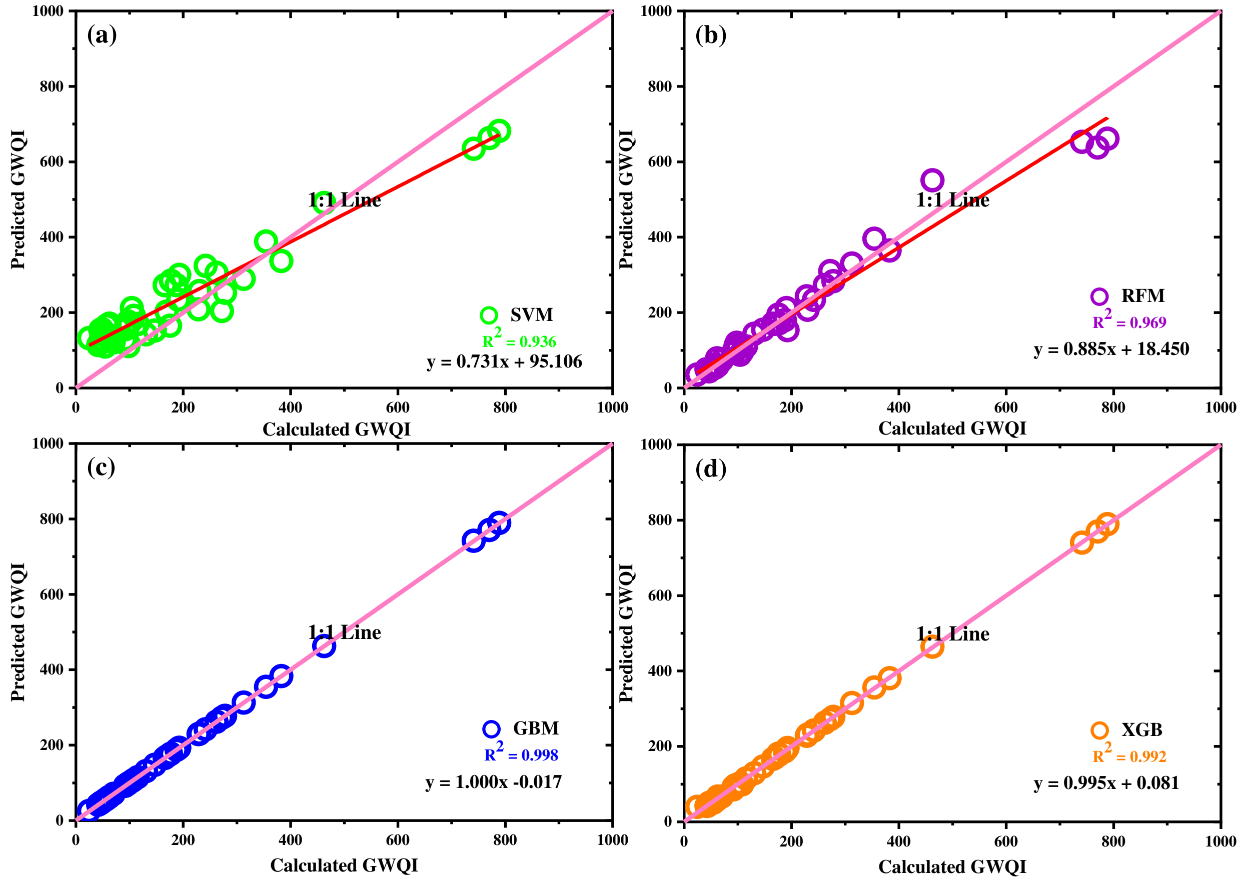

Fig. 5 Scatter plots of the observed and predicted GWQI values by the SVM, RFM, GBM, and XGB models for the training dataset.

Table 6 Model performances for the training and testing datasets

Machine learning models	PBIAS	MAE	RMSE	WI	NSE	R^2
	Training					
SVM	-24.742	68.103	75.884	0.945	0.830	0.936
RFM	1.439	19.023	35.727	0.989	0.962	0.969
GBM	**0.003**	**0.079**	**0.097**	**0.999**	**0.998**	**0.998**
XGB	0.000	2.029	3.260	0.999	0.995	0.992
	Testing					
SVM	19.023	65.165	170.118	0.852	0.651	0.768
RFM	**12.024**	**45.912**	**111.463**	**0.947**	**0.850**	**0.938**
GBM	-7.994	82.169	116.380	0.957	0.836	0.845
XGB	3.552	60.374	126.254	0.941	0.808	0.810

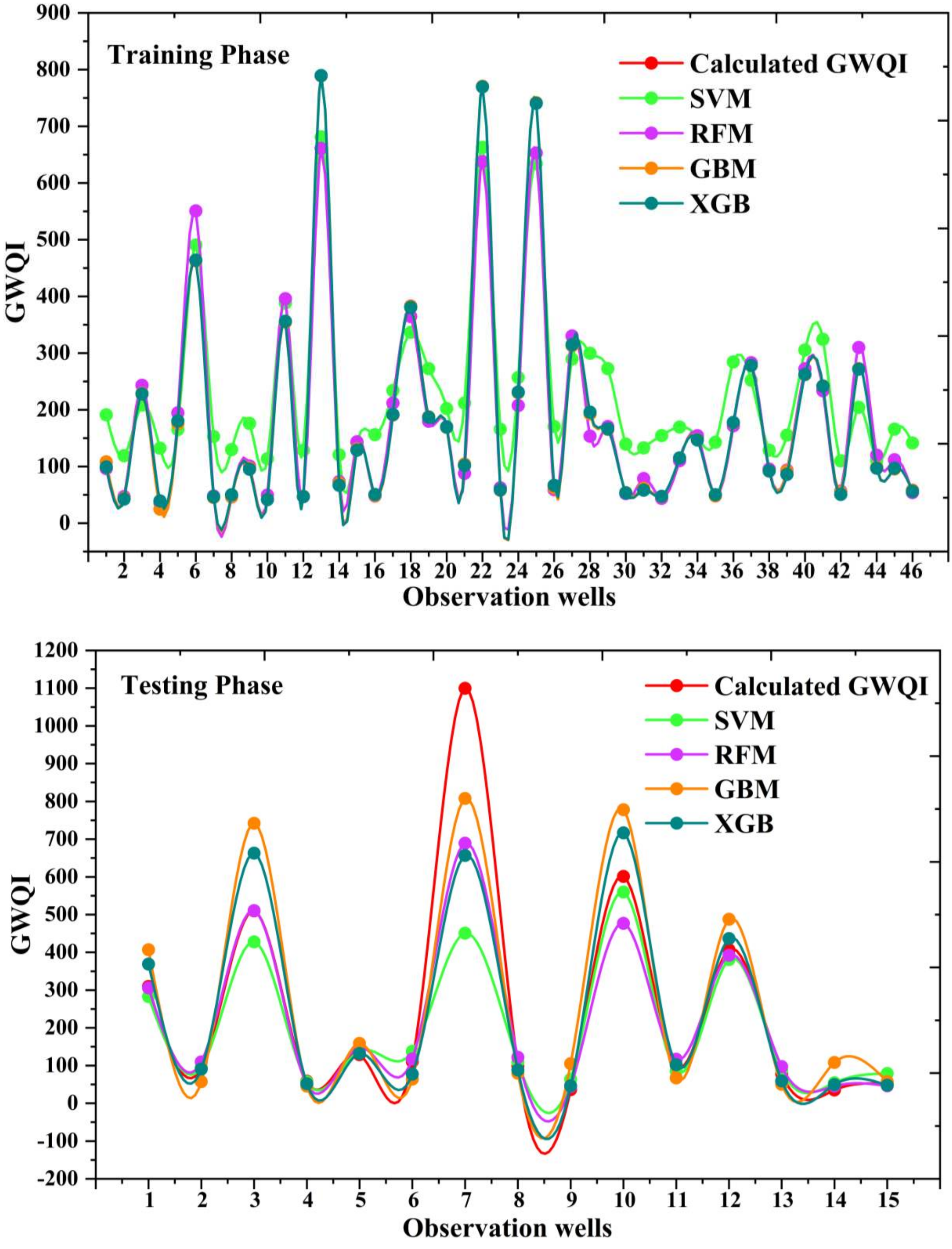

Fig. 6 Line plots between computed and predicted GWQI values by SVM, RFM, GBM, and XGB for the training and testing datasets

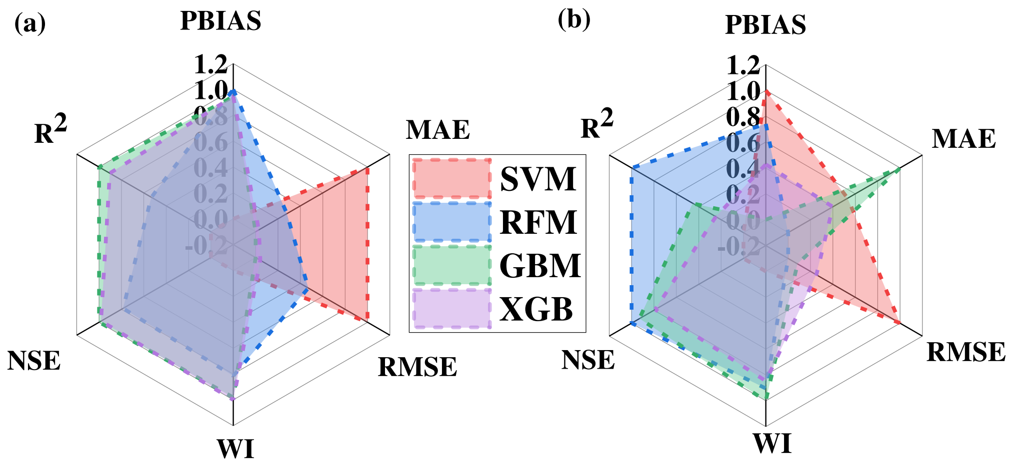

Fig. 7 Radar chart of statistical measures for comparing the model performance (a) training (b) testing

In addition to the above, the comparative analysis of models was done using the Taylor diagram [71] (Fig. 8 a). The SVM model was located furthest. Both the models XGB and GBM were very close to the observed point depending on the standard deviation, correlation, and RMSE. This again showed that the model XGB and GBM competed with each other on the prediction of GWQI. The SVM model also shows significantly worse performances for predicting the GWQI during the training phase. Importantly, during the training process, it was observed that there was no significant superiority observed between the models [85]. However, the validation process, generalization ability evaluation, sensitivity, and uncertainty analysis are important issues in the application of ML in groundwater resource planning and management.

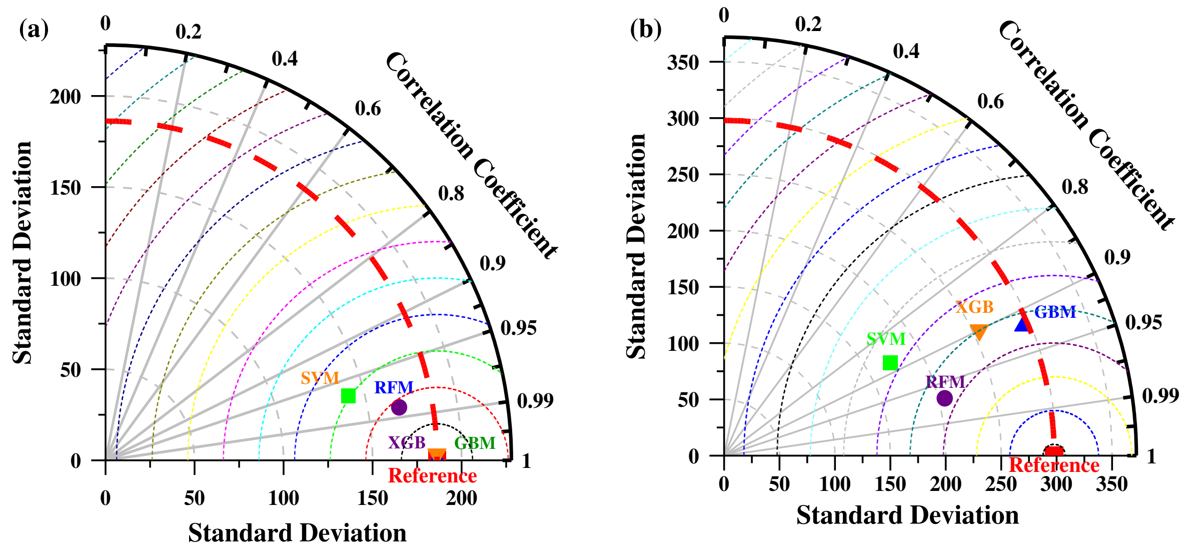

Fig. 8 Graphical comparison of developed models using Taylor diagrams for (a) training (b) testing data sets

3.2.2. Validation of applied ML Models

The trained models were validated using statistical performance criteria i.e., PBIAS, MAE. RMSE, WI, NSE and R^2. Table 6 presents the validation result of applied models for the prediction of GWQI for NCT Delhi. The RFM model showed superiority over other applied models with the statistical indicators, WI (0.850), NSE (0.947) and R^2 (0.938); the lower values of MAE (45.912) and RMSE (111.436) in the prediction of GWQI. It was followed by GBM model (R^2 = 0.845, WI = 0.957, NSE = 0.836, MAE = 82.169, and RMSE = 116.380) and XGB model (R^2 = 0.810, WI = 941, NSE = 0.808, MAE = 60.374, and RMSE = 126.254). The SVM model still has unacceptable performances for simulating GWQI with high values of MAE (65.165) and RMSE (170.118); low values of coefficient of determination (R^2 = 0.768), NSE (0.651) and WI (0.852). For better visualization scatter plots (Fig. 9) were prepared. In scatter plots, the regression line provides the R^2 value as 0.768 for the SVM model, 0.938 for the RFM model, 0.845 for the GBM, 0.631 and 0.810 for the XGB model during the testing stage, respectively. It revealed that predicted values by the RFM model are closely distributed over the 1:1 line better than those of the SVM and, XGB, GBM models. This showed the relatively better performance of the RFM model to other developed models during the validation phase.

Furthermore, the comparison among the developed models was carried out using a series line plot between the computed and predicted values of GWQI (Fig. 6). The line representing the predicted values by the RFM model closely aligns with the line of the calculated GWQI. A detailed comparison was also conducted using a radar chart (Fig. 7b). The RFM model demonstrated the best statistical measures, reflecting its superiority over the other models.

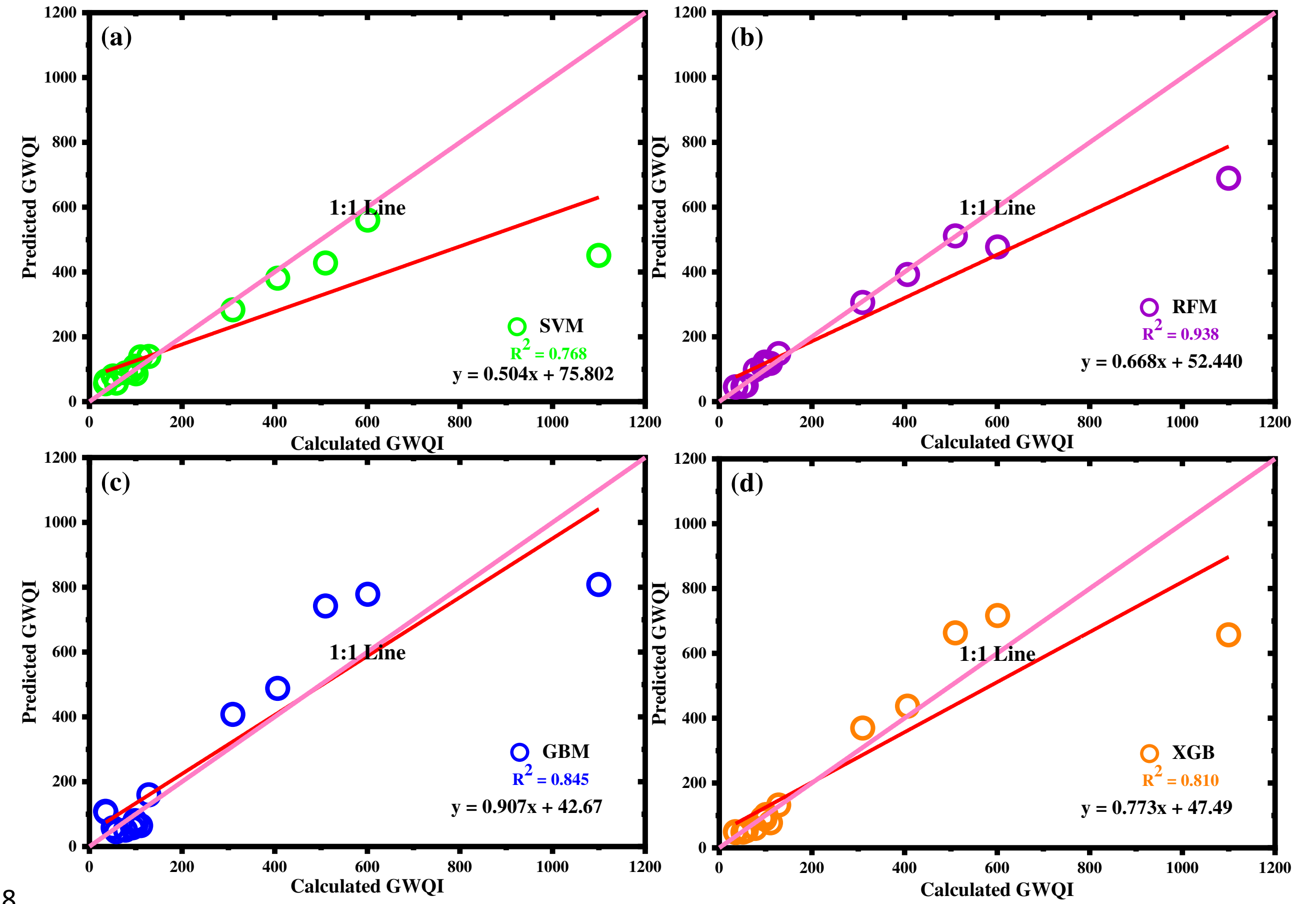

Fig. 9 Scatter plots of the observed and predicted GWQI values by the SVM, RFM, GBM, and XGB models for the testing data samples.

Apart from the above, the Taylor diagram is used to visualise the efficacy of developed models. The fundamental advantage of this graphical approach is that it summarizes three important statistical criteria in a single chart: RMSE, R, and standard deviation (SD) [86]. Furthermore, it displays the model's correctness and realism when compared to the observable parameters. The SD stands for the number of average measurements that deviate from one another. As a result, high precision is indicated by the relative value of standard deviation predicted (SDP) to standard deviation actual (SDA). In contrast, the value of SDP compared to SDA denotes inferior accuracy. From Fig. 8(b), the RFM model is relatively close to the observed point and the SVM model is located farthest from the observed point. This demonstrates the superiority of the RFM model in the prediction of GWQI compared to the SVM, XGB and GBM. However, models RFM and XGM produced acceptable results and the SVM model failed to predict the GWQI during the validation phase as it has a high value of RMSE and low coefficient of determination (R^2) and SD. Fig. 10 presents boxplots illustrating the distribution of estimation errors for GWQI across the models on the testing datasets. These boxplots provide a visual comparison of the variability, central tendency, and outliers in the prediction errors, highlighting the performance and consistency of each model. The boxplot for RFM has a smaller interquartile range (IQR) compared to the others, indicating less variability in prediction errors. Additionally, the RFM model has relatively fewer extreme outliers compared to SVM and XGB, reflecting better consistency. The median value for RFM is closer to the central range, suggesting more accurate predictions. Thus, the RFM model demonstrates superior accuracy and robustness among the models.

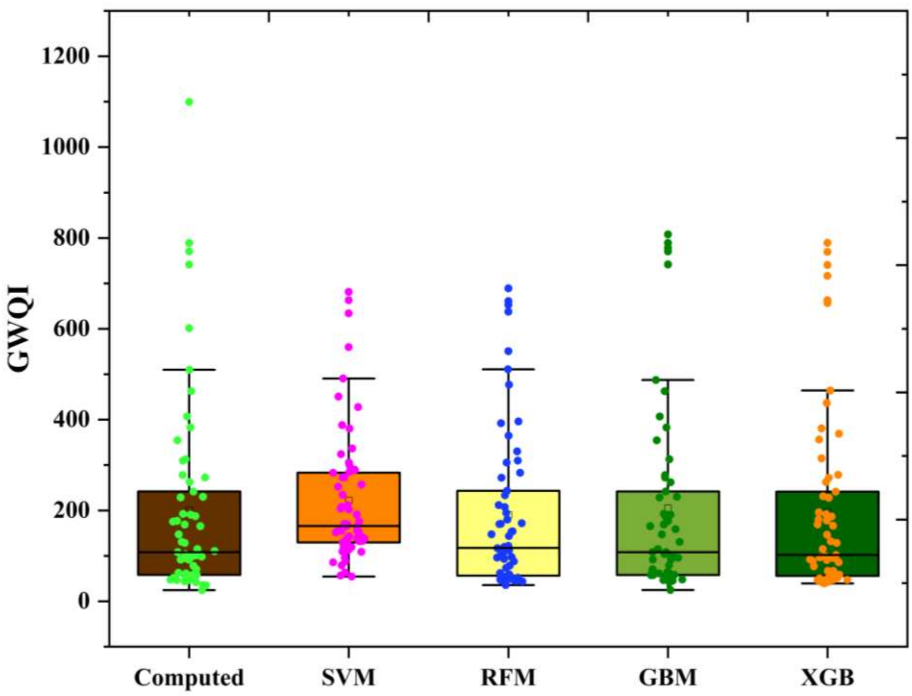

Fig. 10. Boxplots illustrating the models for the testing datsets GWQI estimation errors distribution.

4. Discussion

4.1 Discussion on Machine Learning based GWQI Prediction

The results of the present study demonstrate the efficacy of machine learning (ML) models in predicting GWQI based on a set of influential parameters. In the training phase, the models, particularly Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGB), exhibited remarkable performance, as evidenced by high values of WI, NSE, and R^2, along with low MAE and RMSE. This signifies their robustness in capturing the complex relationships among the groundwater quality parameters. The comparison of observed and predicted GWQI values through scatter plots and line plots further supported the accuracy of GBM and XGB models, emphasizing their ability to align closely with the computed GWQI. The Taylor diagram provided a comprehensive view of model performance, with GBM and XGB models demonstrating close proximity to the observed point. In contrast, the Support Vector Machine (SVM) model exhibited inferior performance during the training phase, underscoring the importance of selecting appropriate ML models for groundwater quality prediction.

Moving to the validation phase, the RFM emerged as the superior model, showcasing higher values of WI, NSE, and R^2, coupled with lower, PBIAS, MAE and RMSE. This emphasizes the ability of developed to generalize well to unseen data, a crucial aspect in the practical application of predictive models. The scatter plots and line plots during the testing stage further confirmed the relatively better performance of RFM compared to other applied models. The Taylor diagram in the validation phase reiterated the dominance of the RFM model. The exceptional performance of the RFM during the validation phase can be attributed to its ensemble learning nature, which harnesses the power of multiple decision trees to collectively enhance predictive accuracy. Groundwater quality prediction inherently involves intricate non-linear relationships, and RFM excels in capturing these complexities, making it particularly suitable for such environmental datasets [26, 87, 88]. The ability of model to determine feature importance facilitates focused analysis, identifying the most influential groundwater quality parameters. The robustness of RFM to noisy data and outliers, common in environmental datasets, ensures stability in the face of real-world data variations.

Moreover, RFM adeptly handles missing data without requiring imputation, a critical advantage when dealing with incomplete groundwater datasets. Its reduced sensitivity to hyperparameter tuning simplifies the model development process, contributing to a balance between bias and variance that prevents overfitting. In groundwater quality prediction scenarios, where datasets may encompass a large feature space, effectiveness of RFM in managing complex data structures is paramount. Additionally, the RFM model ease of implementation and relatively simple hyperparameter requirements facilitate practical application and deployment. While RFM emerged as the top performer in the present study, the contextual appropriateness of a model choice cannot be overstated, as dataset characteristics, problem nature, and study goals should guide the selection of the most suitable machine learning model. The collective strengths of RFM, including ensemble learning, non-linearity handling, feature importance determination, robustness to noise, and simplicity in implementation, collectively contribute to its outstanding performance during the validation phase, underscoring its potential as a robust tool for groundwater quality prediction and management.

The findings from the present study are analogues to Mohseni et al. [2] and Shams et al. [89] where both studies explored the efficacy of ML models and concluded that metaheuristic approaches such as GBM, RFM, XGB provide reliable predictions for water quality assessment. Mohseni et al [2] conducted study to predict the urban water quality index (WQI) for Ujjain city, Madhya Pradesh, India, using four machine learning models (ANN, SVM, RF, and XGB) along with multiple linear regression (MLR). Among these models, XGB outperformed others, achieving the highest accuracy with $R^2 = 0.987$, RMSE = 3.273, and MAE = 2.727 during testing, and an AUC of 0.9048 validated its robustness. This study highlights the effectiveness of ML models, especially XG-Boost, for reliable WQI predictions, aiding decision-makers in urban water management. Another study Shams et al. [89] concluded that Gradient Boosting (GB) achieved 99.50% accuracy for water quality classification (WQC) prediction, while MLP regressor excelled in WQI prediction with $R^2 = 99.8\%$ using a dataset of 7 features. Preprocessing and grid search optimization improved performance, highlighting ML effectiveness for water quality assessment. Similarly, Mo et al. [29] applied XGB and RFM for WQI prediction. They concluded that both models provided the most accurate WQI predictions, especially in winter, using minimal key parameters like Ammonia Nitrogen, Total Phosphorus, Dissolved Oxygen, and turbidity. Accuracy exceeded 80% for good grade predictions in spring and winter but dropped to 70% in summer and autumn. Seasonal variations highlighted worsening nutrient concentrations at coastal stations, emphasizing the need for reliable models in water quality management. Ganga Devi [25] and Sakaa et al. [26] found that RFM has the potential ability to predict the water quality index (WQI). The results of the present study also revealed that the RFM model outperformed SVM, XGB, and GBM in predicting groundwater quality for the NCT, Delhi.

Future research could emphasize the integration of real-time monitoring data with advanced modeling techniques to enhance the accuracy and adaptability of predictions in dynamic environmental conditions. Combining multiple machine learning algorithms, such as ensemble methods or hybrid models, holds the potential to improve the robustness and reliability of groundwater vulnerability assessments. In the present study, the machine learning models were exclusively based on water quality data, without incorporating geological factors that significantly influence groundwater quality dynamics. Geological characteristics, such as permeability, porosity, and mineral composition, play a crucial role in determining aquifer behavior, contaminant attenuation, and groundwater flow patterns. While relying solely on water quality data provided a focused and practical approach for predicting groundwater quality in the urbanized context of NCT Delhi, the exclusion of geological parameters may have limited the models' ability to account for region-specific subsurface processes and constrained their applicability to areas with differing geological conditions. Future studies integrating geological data alongside water quality information could significantly enhance the robustness, reliability, and generalizability of the models, making them more applicable to diverse environmental and geographical settings.

4.2 Spatial Variability of Ground Water Quality in NCT Delhi

The spatial variability of computed GWQI and predicted by the SVM, RFM, XGB and GBM models within the NCT Delhi are presented in Fig. 11. It indicates the different types of water including excellent to good, poor to very poor and unsuitable water in the aquifers. The groundwater quality of the southwestern part of Delhi has been degraded and has poor to unusable water for drinking. This area includes Najafgarh, Dwarka Sec-12, Jharonda Kalam, Tikri Kalal, Ojwah, Jhuljhuli, Vikaspuri, Tagore Garden, Hirankunda, Sainik Vihar locations in NCT Delhi. Most of the south-east and eastern part of NCT Delhi, in areas around the Yamuna flood plain and Delhi Quartzite Ridge zones has excellent to good groundwater covering the stations, Palam Singal Camp, Jheel Khoh, Mayur Vihar, Gazipur crossing, Balbir Nagar, Mubarakpur, Humayan Tomb, Lodhi Garden and Birla Mandir. The computed GWQI values range from 24.72 to 1099.73. The GWQI range, type of water, number, and percentage of samples under each category have been given in Table 7. The analysis of groundwater samples reveals critical concerns regarding water quality. It clearly shows that only 1.64% of the groundwater samples fall under the "excellent" category, indicating a very small proportion of high-quality water that is safe for consumption. Approximately 16.39% of the samples are classified as "good," suitable for drinking with minimal treatment. Meanwhile, 14.759% of the samples fall under the "poor" category, requiring significant treatment before use. Additionally, 13.11% of the samples are categorized as "very poor," indicating they are barely suitable for drinking and may pose significant health risks without advanced treatment. Alarmingly, 54.11% of the groundwater samples are classified as "unsuitable" for drinking. This signifies that more than half of the analyzed samples fail to meet safe drinking water standards and require urgent intervention.

The high percentage of unsuitable groundwater samples underscores the critical need for immediate action. Strategies such as enhancing groundwater recharge through artificial means, implementing effective rainwater harvesting systems, and adopting better land and water management practices are essential. Without urgent measures, the availability of safe drinking water will remain a significant challenge, threatening both human health and sustainable development. Groundwater contamination in South-West Delhi is primarily caused by over-extraction, industrial effluents, and untreated sewage disposal. The depletion of groundwater levels due to excessive extraction legal and illegal has led to increased pollutant concentration. Additionally, industrial and domestic waste discharge has contributed to heavy metal accumulation and microbial contamination. A study investigating groundwater replenishment with tertiary-treated water demonstrated improved groundwater quality and water table levels, highlighting the potential of sustainable water management solutions [90]. The appropriate artificial groundwater recharge and rooftop water harvesting should be implemented to augment groundwater recharge in the area.

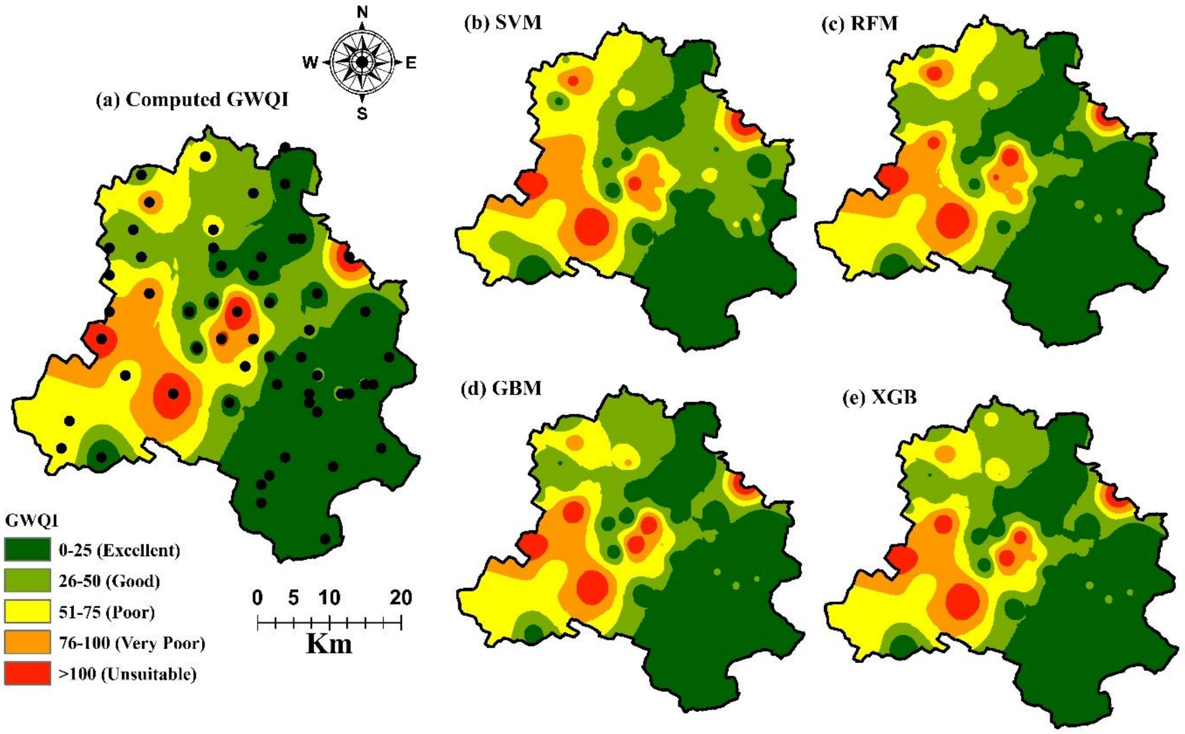

Fig. 11 Spatial variability of GWQI maps (a) computed GWQI; predicted (b) SVM, (c) RFM, (d) GBM and (E) XGB models.

Table 7 The GWQI range, type of water [91, 92], number and percentage of the water samples under each category

Range	Water type	Number of samples	Percentage of the samples
0-25	Excellent water	1	1.64
26-50	Good water	10	16.39
51-75	Poor water	9	14.75
75-100	Very poor water	8	13.11
> 100	Unsuitable	33	54.10

5. Conclusions

Groundwater quality assessment in urban areas is essential for ensuring safe drinking water and protecting public health. It helps identify contamination sources, allowing for timely intervention and mitigation. Regular monitoring also supports sustainable water management by maintaining the balance between supply and demand in urban settings. The present study explored the spatial variability of ground water quality using machine learning approaches in National Capital Territory (NCT) Delhi. The study aims to develop and evaluate four machine learning models, namely Support Vector Machine (SVM), Random Forest Model (RFM), Gradient Boosting Mechanism (GBM), and EXtreme Gradient Boosting (XGB) for modeling ground water quality for 61 sampling locations within the NCT Delhi. Fourteen water quality parameters were used for the computation of ground water quality index (GWQI). Dominance analysis of parameters was performed to select the most influential parameters (i.e., EC, Cl^-, SO_4^{2-}, NO_3^-, Ca^{2+}, Mg^{2+}, Na^+, and TH) for model development and the prediction of GWQI. Results reveled that both the models GBM (R^2 = 0.998) and XGM (R^2 = 0.992) showed superiority during the learning process over the RFM (R^2 = 0.969) and SVM (R^2 = 0.936) models. Although, it was observed that there is no significant superiority between the models. When it comes to the validation, the RFM model with R^2 = 0.938 had better performance than XGB (R^2 = 0.810), GBM (R^2 = 0.845) and SVM (R^2 = 0.768).

It is worth noting that electrical conductivity (EC) is a highly used indicator for water quality determination. The machine learning (ML) models relying on physical parameters as features are efficient tools and should be recommended for forecasting the GWQI for sustainable management of groundwater resources. Our findings on spatial water quality distribution indicated that the groundwater quality of the southwestern part of Delhi has been degraded and has very poor to unsuitable water for drinking. The present study provides information that will help water resource planners to improve groundwater quality by reviving water bodies, sealing illegal bore wells, and constructing water harvesting structures at suitable sides within the NCT Delhi for recharging groundwater. Regularly monitoring groundwater quality is crucial to ensuring its suitability for drinking purposes. Future research could focus on the integration of real-time monitoring data with advanced modeling approaches to improve the accuracy and adaptability of predictions under dynamic environmental conditions. Data size was one of the limitations of the present study. The fusion of multiple machine learning algorithms, such as ensemble methods or hybrid models, could enhance the robustness and reliability of groundwater vulnerability assessments. Furthermore, the Machine learning models developed in present study were solely based on water quality data and did not incorporate geological factors, which play a significant role in groundwater quality dynamics. Geology influences aquifer characteristics such as permeability, porosity, and mineral composition, which can affect the natural attenuation of contaminants and groundwater flow patterns. While the inclusion of water quality data provided a focused and practical approach for predicting groundwater quality in the urbanized context of NCT Delhi, the absence of geological parameters may limit the ability of models to capture region-specific subsurface processes and their applicability to areas with distinct geological conditions. Integrating geological data in future studies could enhance the robustness and generalizability of the models to diverse environmental and geographical contexts.

Declarations

Ethics approval: All authors comply with the guidelines of the *Water Conservation Science and Engineering.*

Consent to participate: All authors agreed to participate in this study.

Consent to publication: All authors agreed to the publication of this manuscript.

Funding: No funding was received to conducting this study.

Conflicts of interest/Competing interests: The authors declare that they have no conflict of interest.

Availability of data and material: Data will be made available on reasonable request.

Code availability: Not applicable.

Authors Contributions:

Nand Lal Kushwaha: Model Development, Writing-review & editing, Writing-original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization.
Madhumita Sahoo: Model Development, Writing-review & editing, Writing-original draft.
Nilesh Biwalkar: Writing-review & editing, Writing -original draft, Supervision.

References

1. Adimalla N, Taloor AK (2020) Hydrogeochemical investigation of groundwater quality in the hard rock terrain of South India using Geographic Information System (GIS) and groundwater quality index (GWQI) techniques. Groundw Sustain Dev 10:100288. https://doi.org/10.1016/j.gsd.2019.100288

2. Mohseni U, Pande CB, Chandra Pal S, Alshehri F (2024) Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model. Chemosphere 352:141393. https://doi.org/10.1016/j.chemosphere.2024.141393

3. Chandra Pal S, Saha A, Kumar Jaydhar A (2024) Groundwater vulnerability assessment of elevated heavy metal contamination related health hazard in coastal multi-aquifers of Sundarban Biosphere Reserve, India. J Hydrol 637:131353. https://doi.org/10.1016/j.jhydrol.2024.131353

4. Abanyie SK, Apea OB, Abagale SA, et al (2023) Sources and factors influencing groundwater quality and associated health implications: A review. Emerg Contam 9:100207. https://doi.org/10.1016/j.emcon.2023.100207

5. Shrestha S, Pandey VP (2016) Chapter 1 - Groundwater as an Environmental Issue in Asian Cities. In: Shrestha S, Pandey VP, Shivakoti BR, Thatikonda S (eds) Groundwater Environment in Asian Cities. Butterworth-Heinemann, pp 1–13

6. Zanotti C, Rotiroti M, Caschetto M, et al (2022) A cost-effective method for assessing groundwater well vulnerability to anthropogenic and natural pollution in the framework of water safety plans. J Hydrol 613:128473. https://doi.org/10.1016/j.jhydrol.2022.128473

7. Saha A, Pal SC, Chowdhuri I, et al (2022) Effect of hydrogeochemical behavior on groundwater resources in Holocene aquifers of moribund Ganges Delta, India: Infusing data-driven algorithms. Environ Pollut 314:120203. https://doi.org/10.1016/j.envpol.2022.120203

8. Ruidas D, Pal SC, Saha A, et al (2024) Ecosystem richness degradation assessment from elevated hydro-chemical properties of Chilka Lake, India. Hydrol Sci J 69:377–389. https://doi.org/10.1080/02626667.2024.2314655

9. Mohseni U, Patidar N, Pathan AI, et al (2022) An Innovative Approach for Groundwater Quality Assessment with the Integration of Various Water Quality Indexes with GIS and Multivariate Statistical Analysis—a Case of Ujjain City, India. Water Conserv Sci Eng 7:327–349. https://doi.org/10.1007/s41101-022-00145-0

10. Patel A, Kethavath A, Kushwaha NL, et al (2023) Review of artificial intelligence and internet of things technologies in land and water management research during 1991–2021: A bibliometric analysis. Eng Appl Artif Intell 123:106335. https://doi.org/10.1016/j.engappai.2023.106335

11. Sahoo M (2022) Chapter 5 - Evaluation of machine learning-based modeling approaches in groundwater quantity and quality prediction. In: Gupta PK, Yadav B, Himanshu SK (eds) Advances in Remediation Techniques for Polluted Soils and Groundwater. Elsevier, pp 87–103

12. Carleo G, Cirac I, Cranmer K, et al (2019) Machine learning and the physical sciences. Rev Mod Phys 91:045002. https://doi.org/10.1103/RevModPhys.91.045002

13. Saha A, Pal SC (2023) Modelling groundwater vulnerability in a vulnerable deltaic coastal region of Sundarban Biosphere Reserve, India. Environ Geochem Health 46:8. https://doi.org/10.1007/s10653-023-01799-y

14. Abd-Elaty I, Kushwaha NL, Patel A (2023) Novel Hybrid Machine Learning Algorithms for Lakes Evaporation and Power Production using Floating Semitransparent Polymer Solar Cells. Water Resour Manag. https://doi.org/10.1007/s11269-023-03565-2

15. Bedi S, Samal A, Ray C, Snow D (2020) Comparative evaluation of machine learning models for groundwater quality assessment. Environ Monit Assess 192:776. https://doi.org/10.1007/s10661-020-08695-3

16. Kushwaha NL, Rajput J, Suna T, et al (2023) Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection. Ecol Inform 75:102122. https://doi.org/10.1016/j.ecoinf.2023.102122

17. DeSimone LA, Pope JP, Ransom KM (2020) Machine-learning models to map pH and redox conditions in groundwater in a layered aquifer system, Northern Atlantic Coastal Plain, eastern USA. J Hydrol Reg Stud 30:100697. https://doi.org/10.1016/j.ejrh.2020.100697

18. Friedel MJ, Wilson SR, Close ME, et al (2020) Comparison of four learning-based methods for predicting groundwater redox status. J Hydrol 580:124200. https://doi.org/10.1016/j.jhydrol.2019.124200

19. Mamat N, Mohd Razali SF, Hamzah FB (2023) Enhancement of water quality index prediction using support vector machine with sensitivity analysis. Front Environ Sci 10:

20. Tabassum S, Kotnala CB, Masih RK, et al (2023) Performance Analysis of Machine Learning Techniques for Predicting Water Quality Index using Physiochemical Parameters. In: 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS). pp 372–377

21. Goodarzi MR, Niknam ARR, Barzkar A, et al (2023) Water Quality Index Estimations Using Machine Learning Algorithms: A Case Study of Yazd-Ardakan Plain, Iran. Water 15:1876. https://doi.org/10.3390/w15101876

22. Yadav P, Chandra M, Fatima N, et al (2023) Predicting Influent and Effluent Quality Parameters for a UASB-Based Wastewater Treatment Plant in Asia Covering Data Variations during COVID-19: A Machine Learning Approach. Water 15:710. https://doi.org/10.3390/w15040710

23. Khoi DN, Quan NT, Linh DQ, et al (2022) Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam. Water 14:1552. https://doi.org/10.3390/w14101552

24. Bui DT, Khosravi K, Tiefenbacher J, et al (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci Total Environ 721:137612. https://doi.org/10.1016/j.scitotenv.2020.137612

25. Ganga Devi V (2019) Random Forest Advice for Water Quality Prediction in the Regions of Kadapa District. Int J Innov Technol Explor Eng 8:1464–1466

26. Sakaa B, Elbeltagi A, Boudibi S, et al (2022) Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. Environ Sci Pollut Res 29:48491–48508. https://doi.org/10.1007/s11356-022-18644-x

27. Nayan A-A, Kibria MG, Rahman MdO, Saha J (2020) River Water Quality Analysis and Prediction Using GBM. In: 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT). pp 219–224

28. Osman A, Najah Ahmed A, Chow MF, et al (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J 12:1545–1556. https://doi.org/10.1016/j.asej.2020.11.011

29. Mo Y, Xu J, Liu C, et al (2024) Assessment and prediction of Water Quality Index (WQI) by seasonal key water parameters in a coastal city: application of machine learning models. Environ Monit Assess 196:1008. https://doi.org/10.1007/s10661-024-13209-6

30. Saeedi M, Abessi O, Sharifi F, Meraji H (2010) Development of groundwater quality index. Environ Monit Assess 163:327–335. https://doi.org/10.1007/s10661-009-0837-5

31. Vasanthavigar M, Srinivasamoorthy K, Vijayaragavan K, et al (2010) Application of water quality index for groundwater quality assessment: Thirumanimuttar sub-basin, Tamilnadu, India. Environ Monit Assess 171:595–609. https://doi.org/10.1007/s10661-009-1302-1

32. Rajankar PN, Tambekar DH, Wate SR (2011) Groundwater quality and water quality index at Bhandara District. Environ Monit Assess 179:619–625. https://doi.org/10.1007/s10661-010-1767-y

33. Sadat-Noori SM, Ebrahimi K, Liaghat AM (2014) Groundwater quality assessment using the Water Quality Index and GIS in Saveh-Nobaran aquifer, Iran. Environ Earth Sci 71:3827–3843. https://doi.org/10.1007/s12665-013-2770-8

34. Batabyal AK, Chakraborty S (2015) Hydrogeochemistry and Water Quality Index in the Assessment of Groundwater Quality for Drinking Uses. Water Environ Res 87:607–617. https://doi.org/10.2175/106143015X14212658613956

35. Dhar A, Sahoo S, Sahoo M (2015) Identification of groundwater potential zones considering water quality aspect. Environ Earth Sci 74:5663–5675. https://doi.org/10.1007/s12665-015-4580-7

36. Varol S, Davraz A (2015) Evaluation of the groundwater quality with WQI (Water Quality Index) and multivariate analysis: a case study of the Tefenni plain (Burdur/Turkey). Environ Earth Sci 73:1725–1744. https://doi.org/10.1007/s12665-014-3531-z

37. Boateng TK, Opoku F, Acquaah SO, Akoto O (2016) Groundwater quality assessment using statistical approach and water quality index in Ejisu-Juaben Municipality, Ghana. Environ Earth Sci 75:489. https://doi.org/10.1007/s12665-015-5105-0

38. Norouzi H, Moghaddam AA (2020) Groundwater quality assessment using random forest method based on groundwater quality indices (case study: Miandoab plain aquifer, NW of Iran). Arab J Geosci 13:912. https://doi.org/10.1007/s12517-020-05904-8

39. Fang Y, Zheng T, Zheng X, et al (2020) Assessment of the hydrodynamics role for groundwater quality using an integration of GIS, water quality index and multivariate statistical techniques. J Environ Manage 273:111185. https://doi.org/10.1016/j.jenvman.2020.111185

40. Ram A, Tiwari SK, Pandey HK, et al (2021) Groundwater quality assessment using water quality index (WQI) under GIS framework. Appl Water Sci 11:46. https://doi.org/10.1007/s13201-021-01376-7

41. Singha S, Pasupuleti S, Singha SS, et al (2021) Prediction of groundwater quality using efficient machine learning technique. Chemosphere 276:130265. https://doi.org/10.1016/j.chemosphere.2021.130265

42. Raheja H, Goel A, Pal M (2021) Prediction of groundwater quality indices using machine learning algorithms. Water Pract Technol 17:336–351. https://doi.org/10.2166/wpt.2021.120

43. Mozaffari S, Javadi S, Moghaddam HK, Randhir TO (2022) Development of the support vector regression–particle swarm optimization simulation-optimization model for the assessment of a novel groundwater quality index. Water Environ J 36:608–621. https://doi.org/10.1111/wej.12801

44. Dimple D, Rajput J, Al-Ansari N, Elbeltagi A (2022) Predicting Irrigation Water Quality Indices Based on Data-Driven Algorithms: Case Study in Semiarid Environment. J Chem 2022:e4488446. https://doi.org/10.1155/2022/4488446

45. Saha A, Pal SC, Islam ARMT, et al (2024) Hydro-chemical based assessment of groundwater vulnerability in the Holocene multi-aquifers of Ganges delta. Sci Rep 14:1265. https://doi.org/10.1038/s41598-024-51917-8

46. Hosseini SM, Mahjouri N (2014) Developing a fuzzy neural network-based support vector regression (FNN-SVR) for regionalizing nitrate concentration in groundwater. Environ Monit Assess 186:3685–3699. https://doi.org/10.1007/s10661-014-3650-8

47. Alnuwaiser MA, Javed MF, Khan MI, et al (2022) Support vector regression and ANN approach for predicting the ground water quality. J Indian Chem Soc 99:100538. https://doi.org/10.1016/j.jics.2022.100538

48. Arabgol R, Sartaj M, Asghari K (2016) Predicting Nitrate Concentration and Its 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

49. He S, Wu J, Wang D, He X (2022) Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere 290:133388. https://doi.org/10.1016/j.chemosphere.2021.133388

50. Chakrabortty R, Roy P, Chowdhuri I, Pal SC (2021) Groundwater Vulnerability Assessment Using Random Forest Approach in a Water-Stressed Paddy Cultivated Region of West Bengal, India. In: Groundwater Geochemistry. John Wiley & Sons, Ltd, pp 392–410

51. Naghibi SA, Hashemi H, Berndtsson R, Lee S (2020) Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. J Hydrol 589:125197. https://doi.org/10.1016/j.jhydrol.2020.125197

52. Park S, Kim J (2021) The Predictive Capability of a Novel Ensemble Tree-Based Algorithm for Assessing Groundwater Potential. Sustainability 13:2459. https://doi.org/10.3390/su13052459

53. Zhang S, Shi Z, Wang G, et al (2022) Application of the extreme gradient boosting method to quantitatively analyze the mechanism of radon anomalous change in Banglazhang hot spring before the Lijiang Mw 7.0 earthquake. J Hydrol 612:128249. https://doi.org/10.1016/j.jhydrol.2022.128249

54. Gupte PR (2019) GROUNDWATER RESOURCES VS DOMESTIC WATER DEMAND AND SUPPLY - NCT DELHI. Central Ground Water Board West Block 2, Wing 3 (GF), Sector 1, R K Puram, New Delhi 110066, Delhi

55. Singaraja C (2017) Relevance of water quality index for groundwater quality evaluation: Thoothukudi District, Tamil Nadu, India. Appl Water Sci 7:2157–2173. https://doi.org/10.1007/s13201-017-0594-5

56. Kamyshova G, Osipov A, Gataullin S, et al (2022) Artificial Neural Networks and Computer Vision's-Based Phytoindication Systems for Variable Rate Irrigation Improving. IEEE Access 10:8577–8589. https://doi.org/10.1109/ACCESS.2022.3143524

57. Kouadri S, Elbeltagi A, Islam ARMdT, Kateb S (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

58. Xiao L, Zhang Q, Niu C, Wang H (2020) Spatiotemporal Patterns in River Water Quality and Pollution Source Apportionment in the Arid Beichuan River Basin of Northwestern China Using Positive Matrix Factorization Receptor Modeling Techniques. Int J Environ Res Public Health 17:5015. https://doi.org/10.3390/ijerph17145015

59. WHO (2011) Guidelines for Drinking-water Quality- 4th ed. World Health Organization, Geneva

60. Azen R, Budescu DV (2006) Comparing Predictors in Multivariate Regression Models: An Extension of Dominance Analysis. J Educ Behav Stat 31:157–180. https://doi.org/10.3102/10769986031002157

61. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

62. Chervonenkis AYa (2013) Early History of Support Vector Machines. In: Schölkopf B, Luo Z, Vovk V (eds) Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Springer, Berlin, Heidelberg, pp 13–20

63. Li B, Yang G, Wan R, et al (2016) Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. Hydrol Res 47:69–83. https://doi.org/10.2166/nh.2016.264

64. Sahoo M, Kasot A, Dhar A, Kar A (2018) On Predictability of Groundwater Level in Shallow Wells Using Satellite Observations. Water Resour Manag 32:1225–1244. https://doi.org/10.1007/s11269-017-1865-5

65. Yu P-S, Chen S-T, Chang I-F (2006) Support vector regression for real-time flood stage forecasting. J Hydrol 328:704–716. https://doi.org/10.1016/j.jhydrol.2006.01.021

66. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

67. Leong WC, Bahadori A, Zhang J, Ahmad Z (2021) Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). Int J River Basin Manag 19:149–156. https://doi.org/10.1080/15715124.2019.1628030

68. Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

69. Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2/3:18–22

70. Guelman L (2012) Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst Appl 39:3659–3667. https://doi.org/10.1016/j.eswa.2011.09.058

71. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmospheres 106:7183–7192. https://doi.org/10.1029/2000JD900719

72. Elbeltagi A, Di Nunno F, Kushwaha NL, et al (2022) River flow rate prediction in the Des Moines watershed (Iowa, USA): a machine learning approach. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02228-9

73. Kushwaha NL, Rajput J, Elbeltagi A, et al (2021) Data Intelligence Model and Meta-Heuristic Algorithms-Based Pan Evaporation Modelling in Two Different Agro-Climatic Zones: A Case Study from Northern India. Atmosphere 12:1654. https://doi.org/10.3390/atmos12121654

74. Kushwaha NL, Rajput J, Sena DR, et al (2022) Evaluation of Data-driven Hybrid Machine Learning Algorithms for Modelling Daily Reference Evapotranspiration. Atmosphere-Ocean 60:519–540. https://doi.org/10.1080/07055900.2022.2087589

75. Samantaray S, Sahoo A (2024) Groundwater level prediction using an improved ELM model integrated with hybrid particle swarm optimisation and grey wolf optimisation. Groundw Sustain Dev 26:101178. https://doi.org/10.1016/j.gsd.2024.101178

76. Odusanya AE, Schulz K, Biao EI, et al (2021) Evaluating the performance of streamflow simulated by an eco-hydrological model calibrated and validated with global land surface actual evapotranspiration from remote sensing at a catchment scale in West Africa. J Hydrol Reg Stud 37:100893. https://doi.org/10.1016/j.ejrh.2021.100893

77. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22:679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

78. Samantaray S, Sahoo A, Baliarsingh F (2024) Groundwater level prediction using an improved SVR model integrated with hybrid particle swarm optimization and firefly algorithm. Clean Water 1:100003. https://doi.org/10.1016/j.clwat.2024.100003

79. Sahoo A, Parida SS, Samantaray S, Satapathy DP (2024) Daily flow discharge prediction using integrated methodology based on LSTM models: Case study in Brahmani-Baitarani basin. HydroResearch 7:272–284. https://doi.org/10.1016/j.hydres.2024.04.006

80. Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82. https://doi.org/10.3354/cr030079

81. Samantaray S, Sahoo A, Satapathy DP, et al (2024) Suspended sediment load prediction using sparrow search algorithm-based support vector machine model. Sci Rep 14:12889. https://doi.org/10.1038/s41598-024-63490-1

82. Willmott CJ (1981) On the Validation of Models. Phys Geogr 2:184–194. https://doi.org/10.1080/02723646.1981.10642213

83. Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. Biometrika 78:691–692. https://doi.org/10.1093/biomet/78.3.691

84. Samantaray S, Sahoo A, Yaseen ZM, Al-Suwaiyan MS (2025) River discharge prediction based multivariate climatological variables using hybridized long short-term memory with nature inspired algorithm. J Hydrol 649:132453. https://doi.org/10.1016/j.jhydrol.2024.132453

85. El Bilali A, Taleb A, Brouziyne Y (2021) Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agric Water Manag 245:106625. https://doi.org/10.1016/j.agwat.2020.106625

86. Vishwakarma DK, Kuriqi A, Abed SA, et al (2023) Forecasting of stage-discharge in a non-perennial river using machine learning with gamma test. Heliyon 9:. https://doi.org/10.1016/j.heliyon.2023.e16290

87. Lap BQ, Phan T-T-H, Nguyen HD, et al (2023) Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. Ecol Inform 74:101991. https://doi.org/10.1016/j.ecoinf.2023.101991

88. Nadiri AA, Barzegar R, Sadeghfam S, Rostami AA (2022) Developing a Data-Fused Water Quality Index Based on Artificial Intelligence Models to Mitigate Conflicts between GQI and GWQI. Water 14:3185. https://doi.org/10.3390/w14193185

89. Shams MY, Elshewey AM, El-kenawy E-SM, et al (2024) Water quality prediction using machine learning models based on grid search method. Multimed Tools Appl 83:35307–35334. https://doi.org/10.1007/s11042-023-16737-4

90. Tyagi RS, Singh SK, Goyal PK (2024) Rejuvenation of water bodies with recycled water. Water Pract Technol 19:839–851. https://doi.org/10.2166/wpt.2024.055

91. Kumar Ravi N, Kumar Jha P, Varma K, et al (2023) Application of water quality index (WQI) and statistical techniques to assess water quality for drinking, irrigation, and industrial purposes of the Ghaghara River, India. Total Environ Res Themes 6:100049. https://doi.org/10.1016/j.totert.2023.100049

92. Balamurugan P, Kumar PS, Shankar K, et al (2020) NON-CARCINOGENIC RISK ASSESSMENT OF GROUNDWATER IN SOUTHERN PART OF SALEM DISTRICT IN TAMILNADU, INDIA. J Chil Chem Soc 65:4697–4707. https://doi.org/10.4067/S0717-97072020000104697