



OPEN Integrated groundwater quality assessment using geochemical modelling and machine learning approach in Northern India

Raisul Islam¹, Alok Sinha², Athar Hussain³, Kamlesh Deshmukh⁴, Mohammad Usama⁵, Salman Ahmed⁶, Rajesh Kumar Deolia⁷, Jitendra Kumar⁸ & Shahjad Ali⁹✉

Groundwater is an essential resource for global drinking and agricultural practices, but it is increasingly threatened by contamination. A comprehensive study was conducted for groundwater quality at 23 different locations in Kasganj, Uttar Pradesh, India, utilizing state-of-the-art Water Quality Indexing (WQI) and Irrigation Water Quality Indexing (IWQI) techniques. A total of one hundred fifteen groundwater samples were analyzed for twelve water quality aspects: pH, total dissolved solids (TDS), total alkalinity, total hardness, calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^+), potassium (K^+), chloride (Cl^-), bicarbonate (HCO_3^-), Sulphate (SO_4^{2-}), Nitrate (NO_3^-), and fluoride (F^-). The results revealed that the TDS levels were alarmingly high, spanning 252 to 2054 ppm with an average of 942 ppm. Similarly, fluoride levels, ranging from 0.21 to 3.80 ppm (average 1.55 ppm), exceeded the World Health Organization's permissible limit of 1.5 ppm. Strong correlations among fluoride levels, alkalinity, pH, Na^+ , and HCO_3^- point to geochemical interactions causing pollution. Piper diagram analysis divided most samples into Ca–Mg–Cl hydrochemical facies, a classification indicating the dominant ions in the water. Mineral saturation indices indicated dolomite, calcite, and aragonite oversaturation, which means these minerals are present in excess, potentially due to the water's high TDS levels. With WQI scores ranging from 63.64 to 221.18, WQI results were concerning: 60.87% of samples were judged unfit for drinking, and 26.08% were relatively poor. These findings raise serious health concerns for the affected populations. Variations in IWQI indicators—Na%, SAR, MH, and KL ratio—informed irrigation fit for different sites. The use of advanced machine learning models (ANN, RF, XGB) for hydrochemical facies analysis, geochemical modeling, and predictive WQI in the sampled area makes the current study unique. To enhance forecast accuracy and support water management, Machine Learning models (Random Forest (RF), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGB)), were used. The outcomes are indicated by better performance by RF with minimum error values (RMSE: 5.97, MSE: 35.69, MAE: 5.49) and a high R^2 value of 0.951. ANN followed closely with an R^2 of 0.957, while XGB achieved an R^2 of 0.831. The performance by RF was the best in WQI prediction among the models tested. The results reveal critical groundwater pollution in the Kasganj area, emphasizing the immediate requirement of focused remedial action and effective water management plans.

Keywords Gibbs diagram, Ganga basin, Water quality indexing, Irrigation water quality indexing, Machine learning, Piper diagram

¹Department of Civil Engineering, GLA University, Mathura, India. ²Department of Environmental Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, India. ³Department of Civil Engineering, Netaji Subhas University of Technology, New Delhi, India. ⁴Department of Computer Science and Engineering, Anand Engineering College, Agra, India. ⁵Department of Environmental Science, Integral University, Lucknow, India. ⁶Interdisciplinary Department of Remote Sensing and GIS Applications, Aligarh Muslim University, Aligarh, India. ⁷Department of Applied Science (Mathematics), G.L. Bajaj Group of Institutions, Mathura, India. ⁸Department of Mathematics and Computing, Madhav Institute of Technology & Science-DU, Gwalior, M.P., India. ⁹Department of Environmental Science, Sharda University Agra, Keetham, Agra 282007, India. ✉email: alishahzad75@gmail.com

Groundwater is a plentiful asset that millions of people throughout the world rely on for their water to consume. The rising prevalence of contaminants in groundwater makes it all the more important to evaluate its quality for human consumption¹. Groundwater is a vital and significant source of potable and irrigation water in regions that experience arid and semi-arid environments. Pollution of ground water has become a global problem due to the increasing human population and the accompanying fast urbanization and industrialization². Over the past few decades, groundwater quality has fallen due to a disruption in chemical processes caused by an increase in human activity^{3,4}. The level of solids and soluble salts determines the irrigation water quality. Evaluating the level of quality is vital for the long-term usage of these natural assets for crop irrigation². The quantity and quality of groundwater are both negatively affected by changes to an area's local terrain and drainage systems⁵. Evaluating the quality and quantity of groundwater is a crucial factor in establishing its viability for drinking and irrigation purposes^{6–8}. The importance of supplying freshwater for industrial, agro-industrial, and household uses has grown in tandem with the expansion of industrialization. A large portion of groundwater, around 65%, is utilized for human consumption, with a smaller percentage going toward irrigation and domesticated animals using 20% and industrial uses and quarries using 15%^{9–11}. A major global problem now is the gradual deterioration of groundwater quality. The rising scarcity of groundwater poses a health hazard to humans, as billions of people around the world are forced to drink polluted water because there is not sufficient potable water. It is now generally known that the cleanliness of groundwater is a greater concern^{12–15}. Approximately eighty percent of worldwide water-related illnesses are caused by water that is not fit for human consumption. But water-related illnesses are killing millions in a number of African, Asian, and Indian states^{16,17,6}. Hypertension, hypocalcaemia, kidney stones, gastro-renal pain, arterial calcification, thrombosis, and other major human health problems have been linked to pollution such as heavy metals, pesticides, and organic and inorganic pollutants, according to previous studies^{17,18,6}. Usually, the Water Quality Index (WQI) is one of the simplest, comprehensive calculative tools for evaluating water quality^{19–21}. The WQI is calculated using a variety of methods, one of which is the water's mathematical single-scoring number⁶. WQI is a method applied to measure the quality of water. It is generally determined by measuring electrical conductivity (EC), pH, sodium ions (Na^+), chloride ions (Cl^-), and bicarbonate ions (HCO_3^-)^{22,23–26}. The quality of groundwater for irrigation is evaluated by sodium percentage (Na%), sodium absorption ratio (SAR), residual sodium carbonate (RSC), permeability index (PI), chlorine index (KI), and magnesium hazard (MH)^{2,26–28}. The indices used in this research, including the Water Quality Index (WQI) and Irrigation Water Quality Index (IWQI), are well-known and effective tools for simplifying complex water quality data. By integrating multiple physicochemical parameters into a single score, these indices provide a clear and comprehensive assessment of overall water quality for both drinking and agricultural purposes. The study uses these standardized methods to give a clear and comparative picture of the quality of the groundwater. This is important for making good decisions about how to manage and fix the water.

Machine learning algorithms enhance and supplement the Water quality index and evaluation. Numerous studies related to gene expression programming (GEP), support vector machines (SVM), artificial neural networks (ANN), and adaptive neuro-fuzzy inference systems (ANFIS) have been employed to assess water quality characteristics. The Automatic Linear Model (ALM) has been utilized to determine the interconnections and key elements that affect structure behavior in many industries in recent studies. This investigation employs indices and the automatic linear model to assess groundwater and identify contaminated sources⁸. The constructed ANN model in this work, with its precise estimation of the proportion of variance in recorded Water Quality Index values, is resilient. Its exceptional concordance with the testing subset deviations and lowest cross-validation measurements indicates this excellent performance. Moreover, the model shows the best R^2 value and a strong connection between projected and absolute WQI values, reassuring its dependability in water quality prediction. A noteworthy scarcity of studies has been identified on the utilization of XGBoost, ANN, and RF models for the prediction of groundwater WQI despite their widespread use in evaluating groundwater quality. A thorough risk assessment will help us comprehend the non-carcinogenic and carcinogenic health implications of polluted water, and these models show promise. The present state of WQI prediction research is insufficient, but these models' predictive ability for water quality parameters gives hope for the future^{29,30,31}.

The aims of this research are as follows:

(a) To investigate the physicochemical characteristics of groundwater in 23 sites at Kasganj, U.P., India, 115 samples were examined for key variables such as pH, alkalinity, total dissolved solids (TDS), fluoride, and different ionic components. (b) This study aims to determine the suitability of groundwater for potable and irrigation purposes using the Water Quality Index (WQI) and Irrigation Water Quality Index (IWQI), which provide a comprehensive classification of water quality across the sampled sites. (c) We aim to evaluate the efficacy of three machine learning models—Random Forest (RF), Artificial Neural Networks (ANN), and Extreme Gradient Boosting (XGB)—in predicting WQI from physicochemical characteristics. (d) To identify contamination areas of concern and group areas by their risk of contamination, which will provide scientific evidence for urgent water management and cleanup plans in the study area.

Methods and materials

Hydrology of study area

The mean annual rainfall is 722.4 mm. The sub-humid climate has a lovely winter season and hot summers. The mean daily maximum temperature in May is 41 °C, the mean daily minimum is 27 °C, and the maximum temperature can reach over 46 °C. The monsoon, with its rapid drop in day temperatures, is a significant factor in the region's climate. January is the coldest, with a mean daily high of 22 °C and a mean daily low of 8 °C. Groundwater occurs in unconsolidated alluvial sediment pore spaces in the sedimentation zone. The top silty, sandy clay beds with kankar support dugwells where groundwater occurs. Deeper aquifers have semi-confined groundwater. Our research, conducted in the Kasganj area, located in the northern portion of the Etah district, has provided precise data on the water levels. During the pre- and post-monsoon periods, the depth of the water

level ranges from 3.11 to 10.24 mbgl and 2.58 to 9.79 mbgl, with a variation of 0.17 to 1.50 m. The water table height ranges from 168 to 157 m above mean sea level (m.a.s.l.), indicating a southeasterly regional groundwater flow³².

Analysis of water quality parameters

To understand the study, which was carried out from August 2023 to July 2024 and is shown on the map in Fig. 1, researchers collected underground water samples from 23 neighboring sites with contaminated water. The study involved the systematic collection of water samples from the tube wells, submersibles, and hand pumps, ensuring that stale water was first evacuated and the samples were then stored in prewashed, high-thickness polypropylene (HDPP) bottles in accordance with standard protocols across various locations in the study area. The analytical methods, including advanced techniques such as titration to measure alkalinity, hardness, and chloride, Ca^{2+} , and Mg^{2+} concentrations, were used. A multi-parameter kit calculated pH and TDS; a flame photometer measured Na^{+} and K^{+} concentrations. Finally, a Shimadzu UV-1800 spectrophotometer analyzed nitrate, sulfate, and fluoride. This advanced method resulted in comprehensive groundwater chemical characterization, therefore providing the scientific quality of the research^{33,34}. The estimated error was less than $\pm 5\%$. The flowchart depicts the methodology of the study region (Figs. 2 and 3).

Geochemical modelling

The PHREEQC geochemical modelling was used to accomplish thermodynamic computations of the *SI* (saturation indices) of the different minerals phases that are common in groundwater (Eq. 1)³⁵.

$$SI = \text{Log} (IAP/K_{sp}) \quad (1)$$

In the above equation, IAP stands for solution ion activity, and since carbonate rocks predominate the aquifer materials in the study region according to estimates of thermodynamic saturation, carbonates have shown to

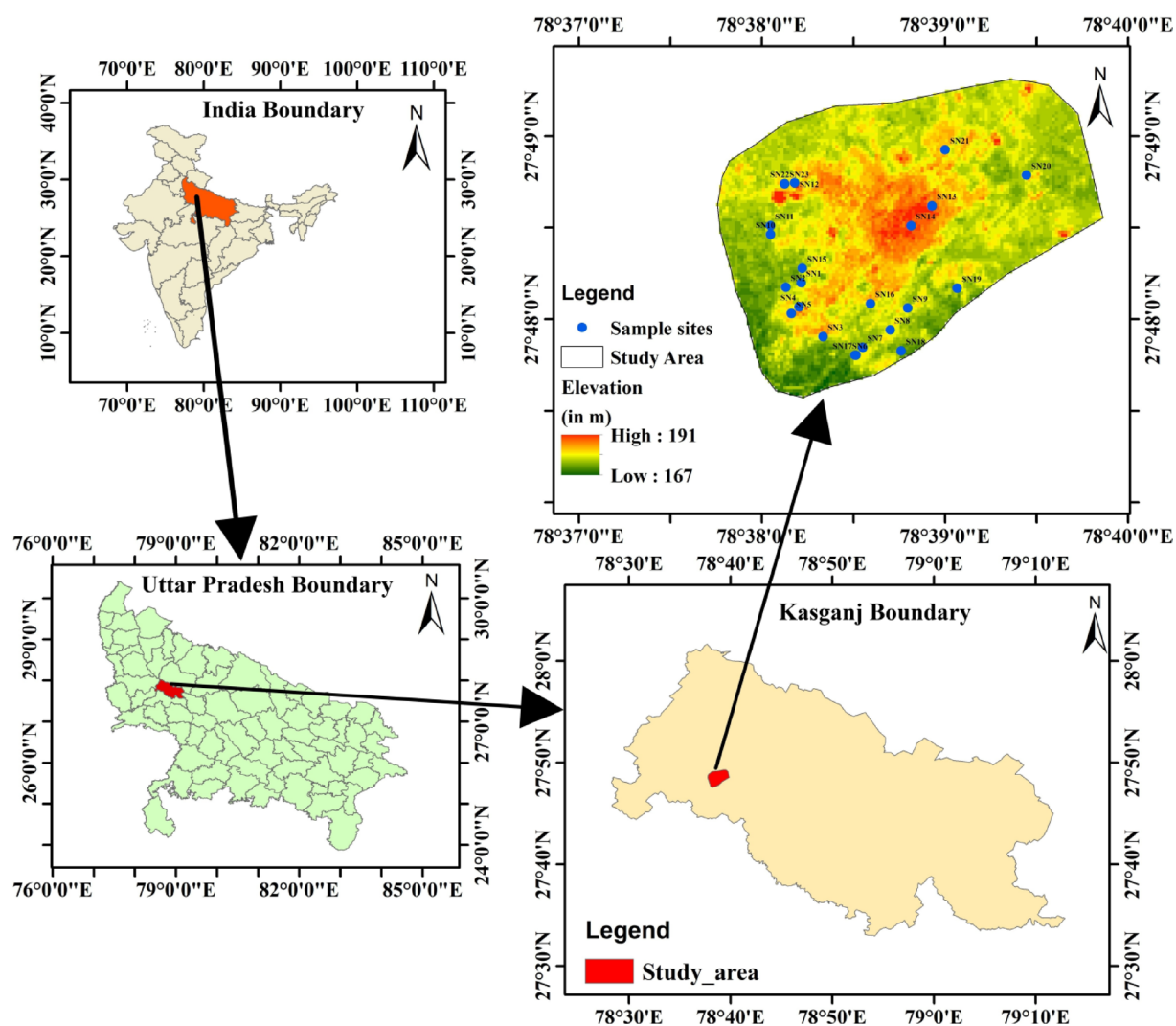


Fig. 1. GIS Location of Kasganj district, Uttar Pradesh, Agra.

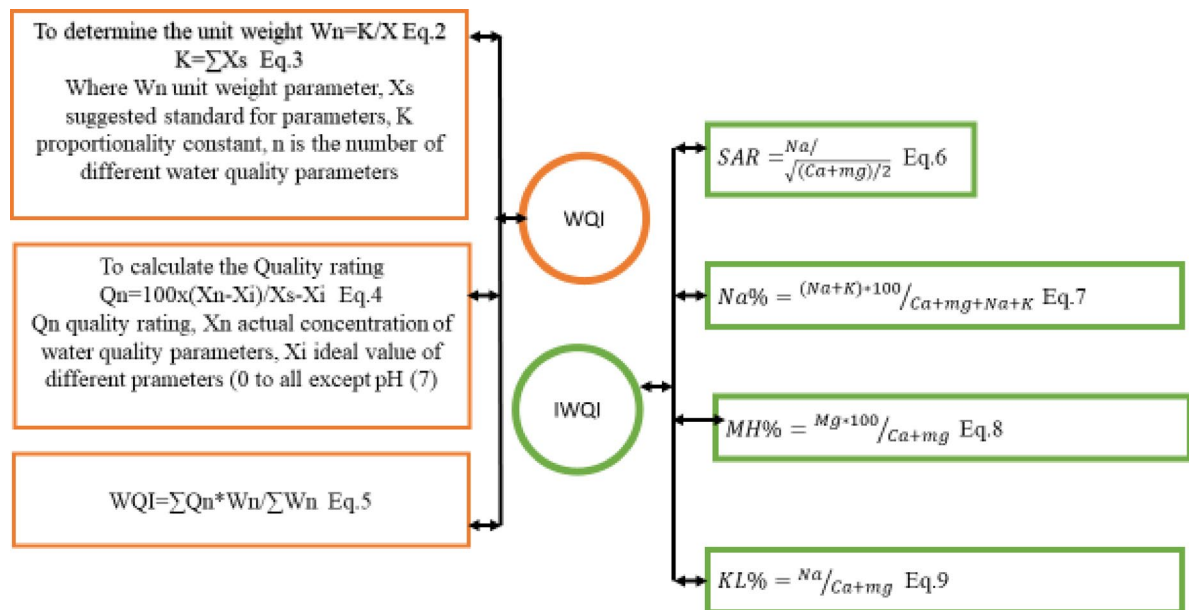


Fig. 2. Flowchart of calculation of water quality indexing (weighted arithmetic index method) and irrigation water quality indexing (SAR, Na%, MH and KR)^{39–41,6}.

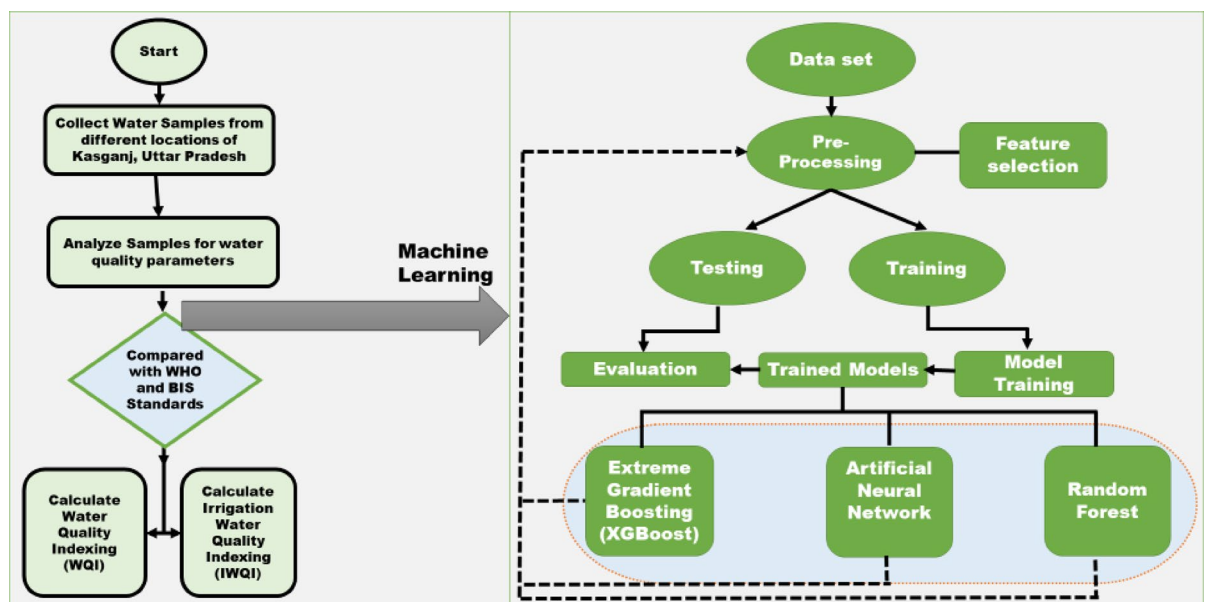


Fig. 3. Flow chart of the methodology implemented for WQI and IWQI analysis of water samples collected from several locations of Kasganj area.

be the most significant minerals in this investigation, K_{sp} is the solubility constant at a given temperature. A groundwater system in an aquifer system, where there is a little amount of the mineral in solution, is represented by a SI level below zero, which indicates that the groundwater is under-saturated with respect to the specific mineral. It also suggests that groundwater has shorter residence spans³⁶. When the saturation index value is greater than zero, it means that the groundwater has reached complete saturation in reaction to the particular mineral present in the solution, meaning that the water can no longer dissolve the mineral.

Estimation of the water quality index (WQI) and irrigational water quality index (IWQI) of the samples

The guidelines given by WHO and BIS standard for drinking water are illustrated in (Table 1) TDS means Total dissolve solids; Na^+ is for Sodium; K^+ stands for Potassium; TH refers for Total hardness; Ca^{2+} stands for Calcium; Mg^{2+} means for Magnesium; TA refers for Total Alkalinity; pH (unitless); F^- term for Fluoride;

Parameter (ppm)	WHO standards	IS 10,500 standards	Unit weight
TDS	1000	500–2,000	0.001
Na ⁺	200	–	0.005
K ⁺	12	–	0.078
TH	500	200–600	0.019
Ca ²⁺	200	75–200	0.005
Mg ²⁺	50	30–100	0.019
TA	500	200–600	0.002
pH (unitless)	8.5	6.5–8.5	0.109
F [–]	1.5	1–1.5	0.62
Cl [–]	250	250–1,000	0.004
NO ₃ [–]	50	45	0.019
SO ₄ ^{2–}	250	200–400	0.004

Table 1. Prescribed water quality and unit weight standards^{37,38}.

WQI Range	Water quality (Type)
> 100	Unfit
76–100	Very poor
51–75	Moderately poor
26–50	Good
0–25	Excellent

Table 2. A classification of drinking water according to the Water Quality Index ranges^{6,39–41}.

S. No	Variables	Range	Water type	References
1	Na%	> 60	Unfit	Wilcox (1955)
		< 60	Suitable	
2	SAR	> 26	Unfit	Reichards (1954)
		18–26	Doubtful	
		10–18	Good	
		< 10	Excellent	
3	MH	> 50	Unfit	Szabolcs and Darab (1964)
		< 50	Suitable	
4	KL	> 1	Unfit	Kelley (1940)
		< 1	Suitable	

Table 3. The groundwater samples classification for irrigation purposes by ranges of Na%, SAR, MH and KL^{27,25,2}.

Cl[–] stands for Chloride; NO₃[–] means Nitrate; (SO₄^{2–}) refers for Sulphate. The estimation of WQI and IWQI are shown in Fig. 2. With this calculation, the water samples have been classified into five different categories of WQI as shown in Table 2

Calculation of the water quality and irrigation water quality indexing (WQI and IWQI)

According to Table 3, IWQI could be classified into five different groups from excellent to unsuitable for irrigation purposes. Based on the results, it was concluded that water available from different sources in this region is not fit for irrigation.

Machine learning models

A wide variety of machine learning classification and prediction techniques have been documented in the literature. Three noteworthy approaches that have demonstrated significant efficacy in a range of applications are examined in this study: Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Random Forest (RF)⁴². Extreme gradient boosting, a novel algorithm gaining popularity for water quality forecasting, is paired with the adaptability of neural networks in handling a large number of inputs and learning nonlinear complex relationships. The three models used in this study are all capable of classification and regression, showcasing their versatility⁴³.

The high-accuracy gradient boost algorithm XGBoost creates a series of decision trees one after the other, allowing every tree to learn and fix the errors of its predecessors. XGBoost broad acceptance is largely due to its strong focus on avoiding overfitting, which maximizes its generalizability. This is made possible by using regularization in input parameters. XGBoost has become a common choice for data science and applied machine learning contexts is great part to its dependability, strong supervised learning algorithm, and efficiency of gradient-boosted models. This method works for regression and classification. Experts recommend XGBoost for its fast execution and out-of-core computation management for small data set. XGBoost has been used in many studies to measure air and water pollution. Gradient-boosted trees combine weak classifiers to form a robust classifier^{44–48}. The boosting process highlights the deficiencies of prior weak classifiers by augmenting the weights or oversampling particular data points. This method instructs the subsequent classifier to concentrate on samples with more significant classification challenges, allowing the model to learn from its prior mistakes. XGBoost, applied in an ensemble learning context, was used to predict regions with elevated lead contamination risk and to determine significant features strongly associated with increased lead levels in Flint, Michigan^{49–52}.

The ANN is the next ML model applied, it composed of interconnected neurons that collaborate to execute particular tasks, taking signals from the biological neural networks observed in nature. The output produced by a neuron arises from a defined process: the neuron takes in input, which is subsequently integrated with coefficients like bias and weights, subsequently, it undergoes processing via a non-linear activation function. Neurons are generally organized in layers, enabling the flow of information from the input layer to the output layer through one or more hidden layers of neurons⁵³. The difference between anticipated and actual results for different input data points is utilized to assess the performance of the network. The loss is utilized to adjust the weights of the network through the application of backpropagation and gradient descent algorithms. This enhances the prediction accuracy and consequently minimizes the losses in subsequent iterations⁵⁴. The essential steps in developing ANN models include selecting suitable inputs and target variable, defining the network's architecture, pre-processing and partitioning the input data, choosing a network design, establishing performance metrics, and performing training, testing, and validation^{31,55–57,5}.

An advanced neural network-based regression model, a significant departure from the traditional non-linear regression model, is employed to accurately predict the Water Quality Index (WQI). This model, which operates on a well-connected parallel with feed-forwarding, is a testament to the innovative strides in our field. The WQI is calculating using F^- , pH, TDS, Cl, Ca, Mg, Na, K, NO_3^- , SO_4^{2-} , TH and TA. The main steps in building this model were choosing the network architecture and structure.

Our model is not only robust and reliable, but also highly adaptable. It uses twelve dimensions (fluoride, pH, TDS, Cl, Ca, Mg, Na, K, NO_3^- , SO_4^{2-} , TH, and TA), hidden layers with various configurations, Rectified Linear Unit (ReLU) activation function, and L2 regularization, K-fold validation, and a 'Linear' output layer targeting WQI. The 5-fold cross-validation tests many hidden layers, learning rates, and regularization strength configurations, showcasing the model's adaptability. Early stopping prevents overfitting, and the second stage chooses optimal training parameters. In parallel with the iteration count, the model is trained using the entire training set and evaluated using the test set. The learning rate and multi-retrain training method ensure the model's robustness and performance^{58,59}.

Random forest makes advantage of an ensemble of classification and regression trees. Every tree is built from the original data set using a distinct bootstrap sample (with replacement) is used to construct each tree from the data set. RF introduces a layer of randomness to the process unlike conventional trees that split each node using the best split among all variables. RF splits a node using just a randomly chosen subset of the variables while building a tree. This fascinating randomness helps, RF resist overfitting in contrast to other techniques. Our model is trained and tested using a large number of trees, which typically improves stability and reduces variance. We have implemented using Random Forest Regressor, hyper-parameters up to 1000 trees with maximum depth to 6, 42 random states, and K-fold cross validation. These techniques efficiently retrain the Model for each fold during cross-validation as a good practice, used to standardized data in each fold, and avoiding data leakage.^{60,61,30}.

The current research work, three ML models (1) RF, (2) ANN, and (3) XGBoost were used to predict and analyze groundwater quality indices. Each model had certain strengths and limitations applicable to groundwater quality evaluation:

Random forest (RF)

Advantages: 1. Resistant to overfitting because of ensemble learning. Handles high-dimensional data and nonlinear relationships well. 2. Offers feature importance for understanding and the Disadvantages: 1. May be computationally expensive with big data. 2. Interpretability is relatively lower compared to linear models.

Artificial neural network (ANN)

Advantages: 1. Able to model intricate, nonlinear interactions between water quality parameters. 2. High predictive accuracy when well-trained and tuned. Disadvantages: (1) Needs great computer power and big data. (2) Functions as a "black box," providing low interpretability of internal processes.

Extreme gradient boosting (XGBoost)

Advantages: (1) High performance and accuracy due to optimized gradient boosting. (2) Effective handling of missing values and overfitting through regularization, and the disadvantage of being more sensitive to parameter tuning.

Variables	Min	Max	Mean	STDEV
F ⁻	0.21	3.80	1.55	1.11
PH	6.99	7.81	7.36	0.20
TDS	252.00	2054.00	942.00	436.19
TA	94.00	456.00	330.13	70.98
Cl ⁻	24.00	290.00	173.13	73.74
Ca ²⁺	24.00	232.00	157.70	52.56
Mg ²⁺	11.60	82.60	54.36	16.87
Na ⁺	20.00	112.00	50.83	28.95
K ⁺	3.00	65.00	17.17	17.53
NO ₃ ⁻	2.80	40.00	17.99	10.98
SO ₄ ²⁻	14.00	121.00	67.21	36.76
TH	108.00	860.00	452.70	168.74
HCO ₃ ⁻	118.80	553.20	400.33	85.25

Table 4. Physical and chemical characteristics (minimum, maximum, mean and standard deviation values) of groundwater samples of groundwater of Ganga Basin Kasganj area, Uttar Pradesh, India.

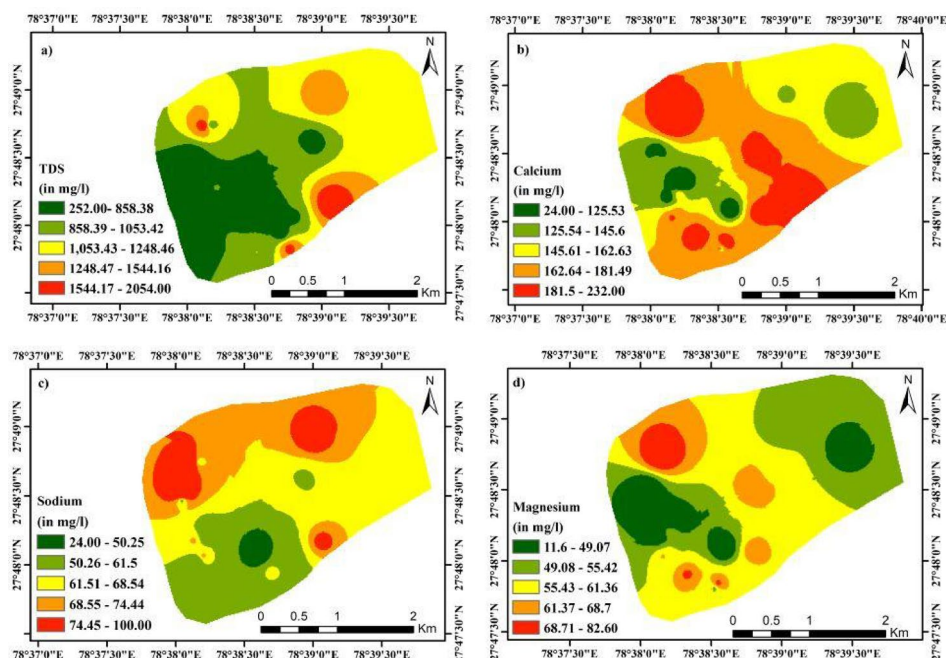


Fig. 5. (a–d) Special distribution of hydrogeochemistry (a) TDS, (b) Calcium (c) Sodium, (d) Magnesium of groundwater of sampled area.

Results

Comprehensive hydrogeochemistry of groundwater of Ganga Basin Kasganj area, Uttar Pradesh, India

Table 4 displays the physico-chemical water quality characteristics of the sampled area. The alkalinity of the water sample was found to be in the range from 94 to 456 ppm. However, the TDS of the samples was alarmingly high, ranging from 252 to 2054 ppm, with an average of 942 ppm (Fig. 5 a–d). The values of chloride, sodium, potassium, sulphate, nitrate, magnesium, calcium ions, and total hardness were within acceptable limits. The mean pH level was 7.36, with a range from 6.99 to 7.81. The concentrations of fluoride in the water samples ranged from 0.21 to 3.80 ppm, with an average of 1.55 ppm, as shown in Fig. 4a–d. These results indicate that the fluoride ion concentration exceeded the World Health Organization acceptable limit of 1.5 ppm (Fig. 5 a–d)³⁷.

Table 5 illustrates the statistical SI values for each mineral in groundwater during the year 2024. Fluorite (CaF₂), Gypsum (CaSO₄), Halite (NaCl), and Sylvite (KCl) were found to be dissolved in the groundwater in mostly wells. The study area is characterized by a shallow aquifer system, exhibiting a transition from unconfined to semi-confined groundwater conditions. The proximity of the water table to the surface has facilitated the formation of clay lenses, which have subsequently introduced an inter-fringing phenomenon within the sandy

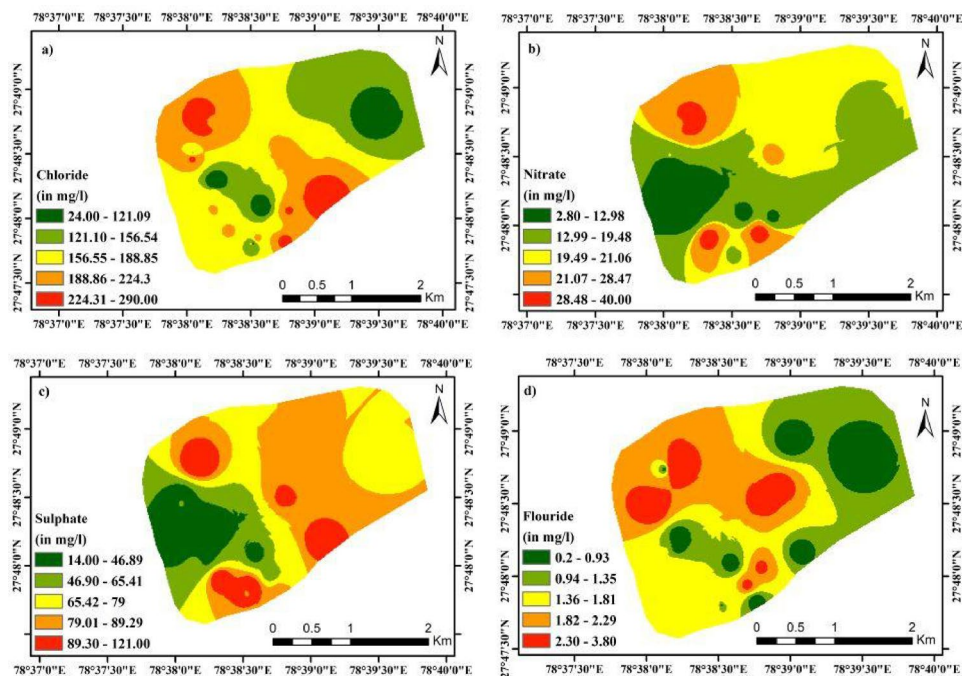


Fig. 4. (a–d) Special distribution of hydrogeochemistry (a) Chloride, (b) Nitrate, (c) Sulphate, (d) Fluoride of groundwater of sampled area.

S No	Anhydrite	Aragonite	Calcite	Dolomite	Fluorite	Gypsum	Halite	Sylvite
SN1	-2.43	0.22	0.36	0.6	-1.26	-2.14	-6.59	-7.05
SN2	-2.63	0.4	0.55	0.97	-0.65	-2.33	-6.47	-7.11
SN3	-1.63	0.49	0.64	1.15	-0.25	-1.34	-6.54	-7.11
SN4	-2.42	0.29	0.43	0.75	-0.29	-2.13	-6.46	-7.03
SN5	-1.89	0.54	0.69	1.24	-0.27	-1.6	-6.62	-7.22
SN6	-1.62	0.26	0.41	0.68	-0.01	-1.33	-6.57	-7.11
SN7	-1.61	0.47	0.62	1.1	-0.25	-1.32	-6.52	-7.16
SN8	-2.03	0.51	0.66	1.19	0.01	-1.74	-6.53	-7.32
SN9	-1.84	0.27	0.41	0.7	0.03	-1.55	-6.51	-7.04
SN10	-2.09	0.81	0.95	1.77	0.17	-1.8	-6.44	-7.31
SN11	-2.55	0.22	0.37	0.61	-0.21	-2.26	-6.4	-7.37
SN12	-1.59	0.73	0.87	1.61	0.42	-1.3	-6.49	-7.17
SN13	-1.78	0.27	0.41	0.7	0.06	-1.49	-6.61	-7.17
SN14	-1.72	0.55	0.7	1.26	0.09	-1.42	-6.46	-7.07
SN15	-2.39	0.32	0.47	0.96	-1.87	-2.1	-7.46	-7.43
SN16	-2.43	0.41	0.56	1.03	-1.61	-2.14	-7.34	-7.05
SN17	-2.02	0.44	0.59	1.07	-2.31	-1.73	-7.05	-6.79
SN18	-1.82	0.52	0.67	1.23	-1.64	-1.53	-6.48	-6.21
SN19	-1.69	0.86	1	1.81	-1.22	-1.4	-6.26	-5.91
SN20	-1.92	0.38	0.52	0.93	-1.68	-1.63	-6.84	-6.65
SN21	-1.85	0.5	0.65	1.2	-1.48	-1.56	-6.6	-6.48
SN22	-1.74	0.44	0.58	1.11	-0.99	-1.45	-6.23	-6.08
SN23	-3.08	-0.9	-0.75	-1.48	-2.46	-2.79	-7.56	-7.31

Table 5. Illustrates the statistical saturation index values for each mineral in groundwater in sampled area.

aquifer, effectively rendering it impermeable. This impermeable layer significantly restricts groundwater recharge. Furthermore, the presence of gravel nodules composed of calcium carbonate within the sandy aquifer influences the pH of the groundwater, thereby enabling the dissolution of minerals. The variance of the saturation index value of a few dissolved minerals in water samples of different wells, such as anhydrite (CaSO_4),

gypsum (CaSO_4), halite (NaCl), sylvite (KCl), was found to be under saturated. The chemical composition of these minerals, which mainly include SO_4 and Cl , shows high value in the study area that is due to anthropogenic contamination (Table 6). SI values that are negative suggest that the water sample exhibits greater aggressiveness towards corrosion.

Table 6 shows the mineralogical analysis of sediment samples from the semi-arid Kasganj area reveals that carbonate minerals—especially Dolomite and Calcite—are the most abundant, as indicated by their relatively high mean values and moderate variability, reflecting favorable alkaline and evaporative conditions typical of such climates. Aragonite also shows stable but lesser presence. In contrast, evaporite minerals like Halite and Sylvite are scarce, displaying very low mean values and limited variability, suggesting that conditions required for their widespread deposition are rare and localized. Sulphate minerals Anhydrite and Gypsum appear in low quantities, possibly due to seasonal hydrological fluctuations that inhibit extensive precipitation. Fluorite exhibits the highest variability, likely linked to local groundwater chemistry differences. Environmentally, this distribution pattern underscores the influence of high evaporation, intermittent water availability, and geochemical processes in shaping mineral assemblages in the region. In conclusion, the data indicates that Kasganj semi-arid environment primarily supports carbonate formation, with evaporite and sulphate minerals occurring only under specialized, occasionally met conditions.

Geochemical characterization of Kasganj, U.P., India

The hydrochemical characterization with the Piper diagram indicates that most groundwater samples adhere to the Ca–Mg–Cl facies (Fig. 6). In the cation triangle, samples predominantly belong to the no dominating type (Field D). Still, they appear to be calcium-rich, indicating mixed cationic contributions from silicate weathering and limited ion exchange processes. The anion triangle exhibits most chloride (Field G), reflecting the influence of evaporite dissolution, anthropogenic causes, or salty water incursion. The center diamond field verifies the categorization inside the Ca–Mg–Cl + SO_4 hydrochemical zone. This zone is often linked with mineralized, hard water and indicates prolonged residence durations or pollution from agricultural and residential activities. The close clustering of sample points implies an incredibly similar hydrogeochemical signature across the research area. Furthermore, the minimal prevalence of bicarbonate-rich facies indicates that recent recharge or carbonate lithology had a limited impact. Overall, the findings provide insight into how a groundwater system is impacted by natural geological processes and perhaps anthropogenic pressures.

By concentrating on the relationship between the concentrations of cations (Na^+ , Ca^{2+}), anions (Cl^- , HCO_3^-), and TDS (Total Dissolved Solid), the Gibbs diagram is a technique for determining the origin of ions in groundwater. To comprehend the relationship between the chemical components of water, the Gibbs diagram was devised (Gibbs 1970, Eqs. (10, 11) Three separate fields of the Gibbs diagram—precipitation dominance, evaporation dominance, and rock–water interaction dominance—are used to identify the quality features of water. All ions are represented in mg/L.

$$\text{Gibbs ratio I (for anion)} = \frac{\text{Cl}^-}{(\text{Cl}^- + \text{HCO}_3^-)} \quad (10)$$

$$\text{Gibbs ratio II (for cation)} = \frac{\text{Na}^+ + \text{K}^+}{(\text{Na}^+ + \text{K}^+ + \text{Ca}^{2+})} \quad (11)$$

Each cation and anion in groundwater has a rock-dominance origin, according to the Gibbs diagram based on TDS and the concentration of cations and anions in Fig. 7. This trait shows that groundwater ion dissolution from interactions with rock or soil is more prevalent than precipitation or other natural sources.

Correlation analysis of Ganga basin area of Kasganj, U.P, Northern India

The present study investigates the correlation of fluoride concentration with other physicochemical characteristics in groundwater samples from the Ganga basin area of Kasganj, Uttar Pradesh, India. Table 7 shows that fluoride has a minimal connection with pH, TA, and HCO_3^- . We discovered a strong positive correlation between fluoride (F^-) ions and bicarbonate (HCO_3^-), sodium (Na^+), and hydrogen (H^+) ions, which is in line with earlier research. This could be because fluoride-containing minerals like fluorite dissolve more

Variable	Minimum	Maximum	Mean	STDEV
Anhydrite	− 3.08	− 1.59	− 2.03	0.40
Aragonite	− 0.9	0.86	0.39	0.33
Calcite	− 0.75	1	0.54	0.33
Dolomite	− 1.48	1.81	0.97	0.63
Fluorite	− 2.46	0.42	− 0.77	0.87
Gypsum	− 2.79	− 1.3	− 1.74	0.40
Halite	− 7.56	− 6.23	− 6.65	0.36
Sylvite	− 7.43	− 5.91	− 6.96	0.42

Table 6. Depicts the minimum, maximum, mean and standard deviation values of saturation index in the sampled area^{62–64}.

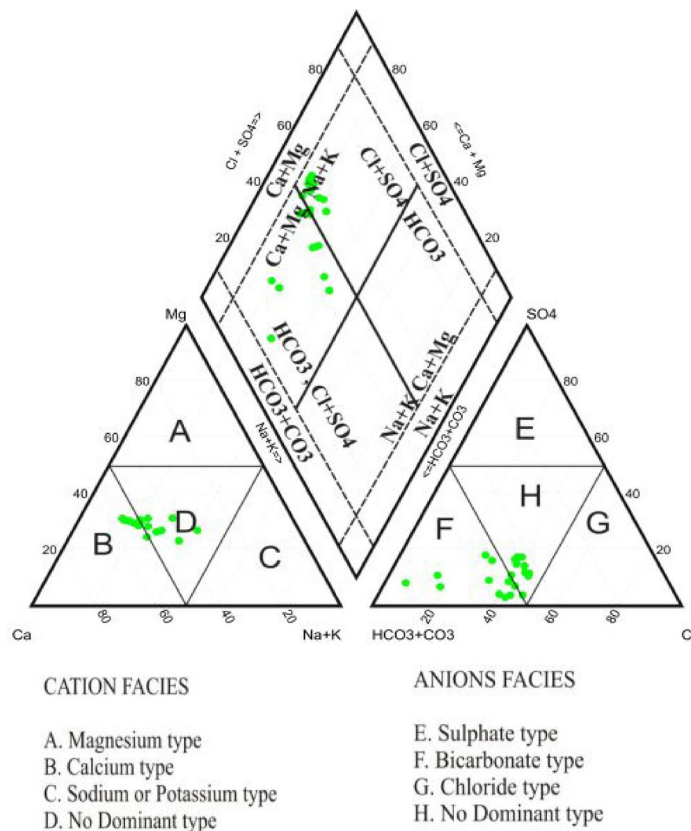


Fig. 6. Piper diagram for samples of groundwater at Kasganj, U.P., India.

readily in alkaline environments ($\text{pH} > 7.5$). However, we discovered a 0.03 correlation value between pH , TA , and HCO_3^- in our instance. Localized geochemical control, such as weathering of calcium and fluoride minerals, or human influence, such as phosphate fertilizers (which include both Ca and F), could be the cause of this. The modest connection between calcium and fluoride (0.48) supports mobilization based on minerals, potentially from complex fluorapatite or mixed silicates instead of pure CaF_2 routes. Low correlations with bicarbonate and pH indicate that local lithology and mineral composition are more important drivers than ion exchange or alkalinity processes. Because our water type is rock dominating, as shown by the Gibbs diagram, fluoride exhibits a significant negative correlation of -0.69 with potassium (K^+). This suggests that fluoride solubility may be influenced by a reverse ion exchange involving Na^+ , Ca^{2+} , and K^+ ^{65,66}.

Spatial distribution of WQI

As illustrated in Fig. 8 a, b, the distribution trend in water quality indexing presents a relatively clear picture. Of the water examined in the Kasganj area, 60.87 percent was deemed unsuitable for human consumption. None were as good, 13.04 percent were classed as moderately poor, and 26.08 percent as extremely poor. Table 8 demonstrates the % distribution of numerous groundwater types in the research geographical area, therefore stressing the serious and alarming character of the problem. The water fluoride and TDS exceed World Health Organization standards and IS10500^{37,38}, indicating a serious issue that demands a swift and effective response and treatment. Experimental results confirmed a high fluoride concentration in water samples of Ganga basin area of Kasganj, U.P, India, which might be due to its geological conditions. It was concluded that water available from different sources in this region is not fit for drinking.

Table 8 comprehensively shows the Kasganj WQI, demonstrating the highest and lowest values across different sampled areas. The range of potential values, a key aspect of our research, is presented, with 63.64 (Saiyad Nagla) representing the water quality index. After a thorough examination, it becomes clear that the maximum value of WQI is in Tarora (221.18).

Table 9 demonstrated that the WQI of the study region, ranging from (233.16, 185.86, 1588, 221.18), reflects water quality influenced by both geogenic and anthropogenic sources, with fluorite rock playing a pivotal role by contributing minerals that significantly impact water chemistry and overall quality.

Irrigation water quality

Irrigation water of low quality may affect crop yields and quality⁶⁸. In the study by⁶⁹, salinity is the primary determinant of irrigation water quality. In the present investigation, we assessed the water's potential for agricultural use by calculating its Na, SAR, MH, and KR percentages.

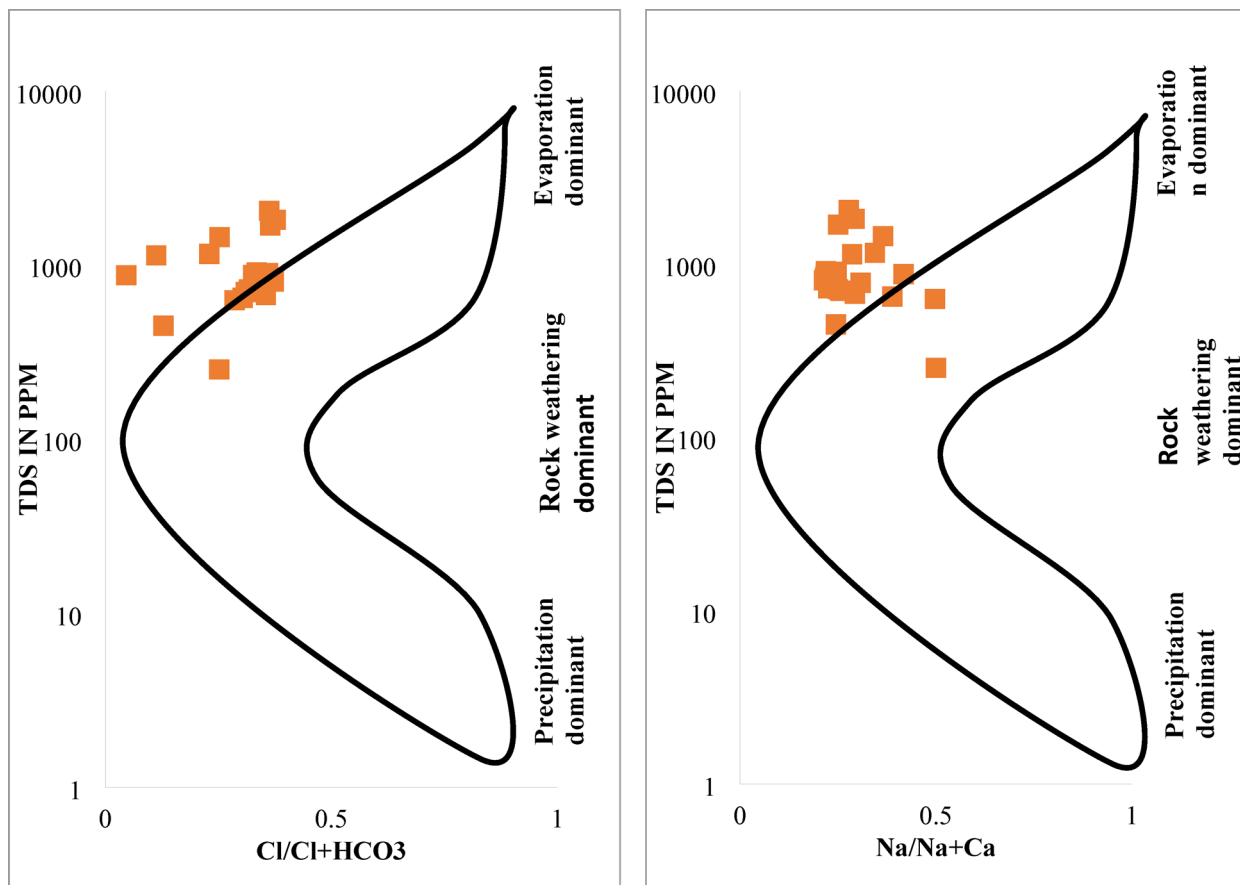


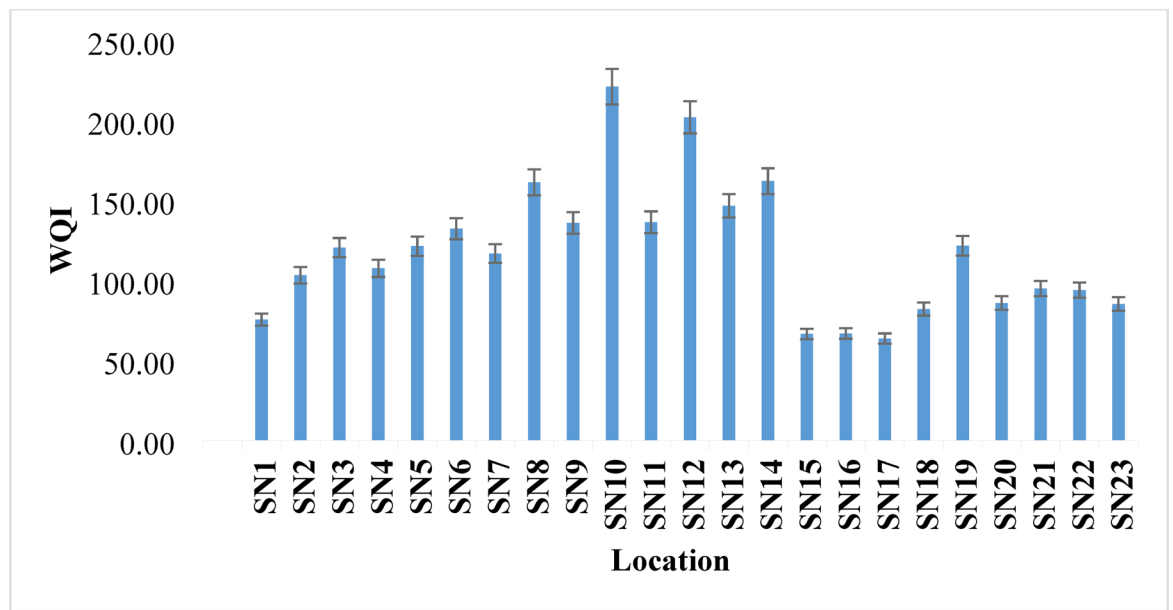
Fig. 7. Gibbs diagram for samples of groundwater at Kasganj, U.P., India.

Variable	F ⁻	PH	TDS	TA	Cl ⁻	Ca ²⁺	Mg ²⁺	Na ⁺	K ⁺	NO ₃ ⁻	SO ₄ ²⁻	TH	HCO ₃ ⁻
F ⁻	1.00												
PH	0.03	1.00											
TDS	-0.33	0.29	1.00										
TA	0.03	0.32	0.46	1.00									
Cl ⁻	0.44	0.00	0.47	0.33	1.00								
Ca ²⁺	0.48	-0.11	0.41	0.32	0.78	1.00							
Mg ²⁺	0.43	-0.16	0.44	0.33	0.73	0.97	1.00						
Na ⁺	0.22	0.12	0.47	0.29	0.47	0.36	0.35	1.00					
K ⁺	-0.69	0.32	0.84	0.23	0.02	-0.05	-0.03	0.19	1.00				
NO ₃ ⁻	0.35	0.13	0.32	0.14	0.34	0.59	0.60	0.39	0.02	1.00			
SO ₄ ²⁻	0.27	0.11	0.53	0.20	0.48	0.79	0.78	0.19	0.20	0.64	1.00		
TH	-0.11	0.09	0.91	0.43	0.61	0.71	0.75	0.43	0.62	0.49	0.73	1.00	
HCO ₃ ⁻	0.03	0.33	0.46	1.00	0.32	0.31	0.33	0.29	0.24	0.14	0.20	0.43	1.00

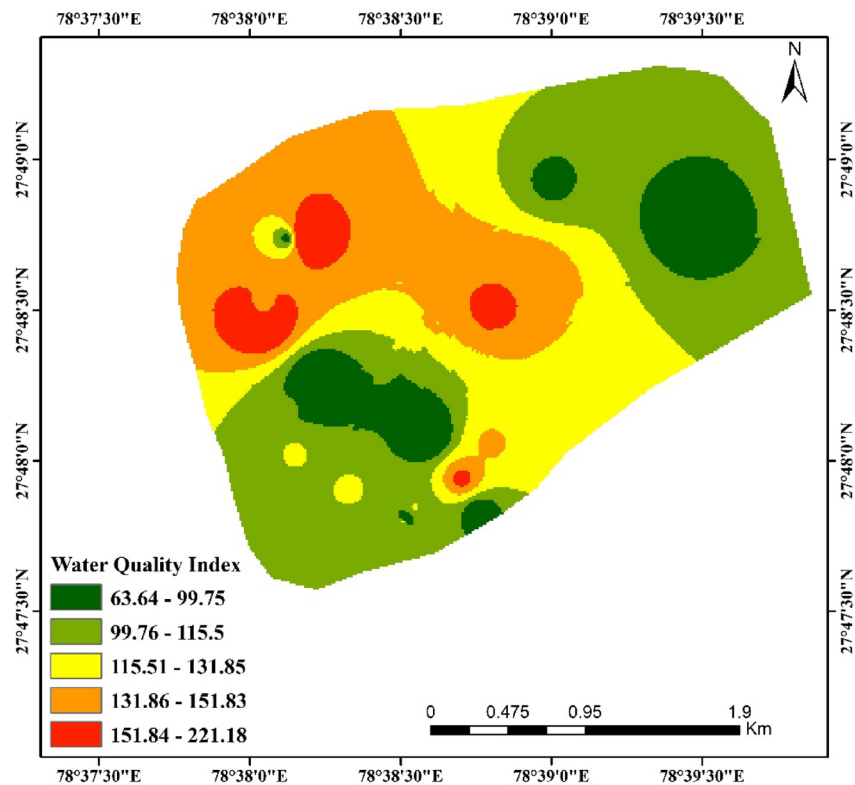
Table 7. Correlation analysis of Ganga basin area of Kasganj, U.P., India.

Percentage of sodium, sodium absorption ratio, MH and KR

The percentage of sodium, SAR, MH, and KL were calculated using Eq. 5, 6, 7, and 8 to determine all the collected samples. The results indicated that the average values Na%, SAR, MH, and KL in the above samples were 25.30%, 7.45%, 26.07%, and 0.29% meq/L, respectively (Figs. 9 and 10a–d). These numbers not only show the suitability of quality of water for irrigation and agricultural uses (Table 10) but also convey possible advantages, such the decrease of soil permeability and the reasons of soil hardness, which could result in better agricultural practices^{68–71}



(a)



(b)

Fig. 8. (a, b) Special graphical and distribution representation of the WQI of the Ganga basin area of Kasganj, U.P, India.

S.NO	WQI	Water quality
SN1	75.46	Very poor
SN2	103.18	Unfit
SN3	120.49	Unfit
SN4	107.52	Unfit
SN5	121.40	Unfit
SN6	132.33	Unfit
SN7	116.75	Unfit
SN8	161.31	Unfit
SN9	135.94	Unfit
SN10	221.18	Unfit
SN11	136.35	Unfit
SN12	202.03	Unfit
SN13	146.57	Unfit
SN14	162.04	Unfit
SN15	66.45	Moderate poor
SN16	66.82	Moderate poor
SN17	63.64	Moderate poor
SN18	82.02	Very poor
SN19	121.65	Unfit
SN20	85.84	Very poor
SN21	94.84	Very poor
SN22	93.89	Very poor
SN23	85.24	Very poor

Table 8. Illustrates the water quality indexing in Kasganj region, North India.

Areas	WQI ranges	References
Barauli Ahir block, Agra district	233.16	Ali et al., 2022
Achhnera block, Agra region, North India	185.86	Ali et al., 2024
Mathura district, Northern India	1588	Salman et al., 2020
Kasganj, Uttar Pradesh, North India	221.18	Present study

Table 9. Illustrated the compared of WQI with previous studies^{1,6,67}.

Application of XGBoost (XGB), artificial neural network (ANN), and random forest (RF), models to predict the quality of water in Kasganj areas

This present research used a data partitioning technique, designating 80% of the dataset for training and 10% each for validation and testing. The predictive performance of three machine learning models—XGBoost (XGB), Artificial Neural Network (ANN), and Random Forest (RF)—in forecasting the Water Quality Index (WQI) over 23 monitored sites is assessed in this work. These discoveries could significantly influence machine learning, environmental science and engineering, and other fields, opening new avenues of exploration. Cross-validation split the dataset with 18 sites set aside for training and 5 for testing every iteration. With R^2 values of 0.9568 for XGB, 0.9994 for ANN, and 0.9368 for RF, the models showed great accuracy throughout the training phase, suggesting strong positive correlations between anticipated and absolute WQI values. The models kept strong performance during the test, producing R^2 values of 0.8427 (XGB), 0.8738 (ANN), and 0.9034 (RF). Via visual regression analysis in Figs. 11a–d, 12a–d, 13a–d, these results confirm the models’ efficacy in WQI prediction; RF shows the best generalizing capacity on unseen data.

Performance of comparative analysis of XGBoost (XGB), artificial neural network (ANN), and random forest (RF), for regression

As detailed in Table 11, the current research assesses the performance of Water Quality Index (WQI) prediction by utilizing three popular machine learning algorithms: XGB, ANN, and RF. Utilizing basic evaluation parameters—Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2)—theoretical foundations, forecasting accuracy, and overall efficacy of each model were rigorously investigated. The outcomes are indicated by better performance by RF with minimum error values (RMSE: 5.97, MSE: 35.69, MAE: 5.49) and a high R^2 value of 0.951. ANN followed closely with an R^2 of 0.957, while XGB achieved an R^2 of 0.831. The performance by RF was the best in WQI prediction among these models tested.

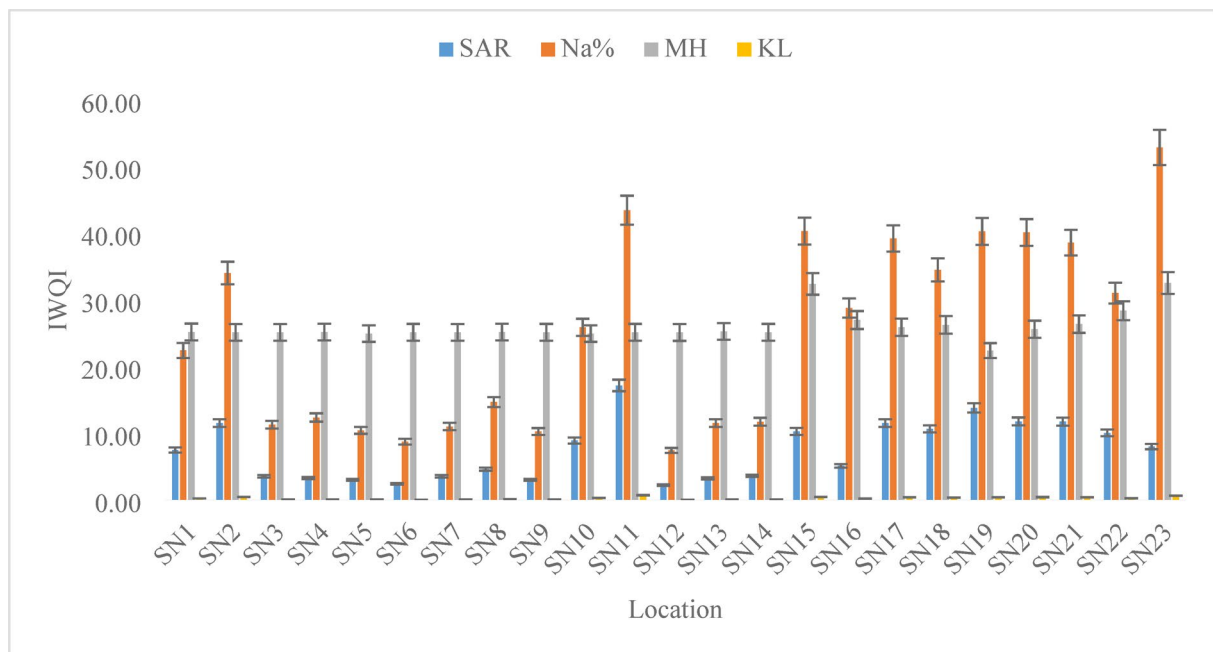


Fig. 9. Special graphical representation of the IWQI of the Ganga basin area of Kasganj, U.P, India.

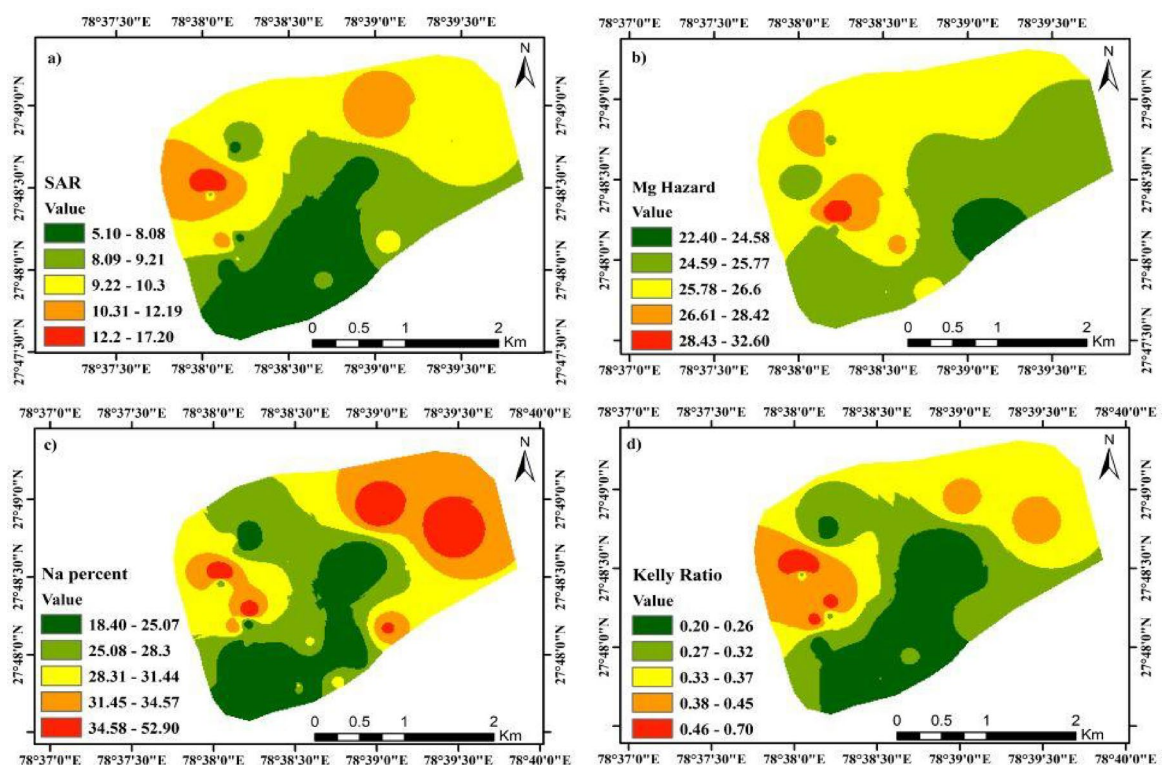


Fig. 10. (a–d) Spatial distribution of IWQI (SAR, Mg Hazard, Na% and Kelly ratio) in the sampled area.

Discussions

The models of machine learning perform well in predicting WQI, with RF showing the highest accuracy. The RF model efficacy for the fluoride and sulfate contaminated of groundwater quality assessment is confirmed by its excellent R^2 value of 0.951 and its low error values. Although other models such as ANN and XGBoost also showed strong performance, RF consistent accuracy across training and validation sets underscores its dependability as a predictive tool. The present research investigation reveals how traditional WQI techniques

S.No	Variables	Range	Water type	Samples %
1	Na%	> 60	Unfit	Nil
		< 60	Suitable	100
2	SAR	> 26	Unfit	Nil
		18–26	Doubtful	Nil
		10–18	Good	39.13
		< 10	Excellent	60.87
3	MH	> 50	Unfit	Nil
		< 50	Suitable	100
4	KR	> 1	Unfit	Nil
		< 1	Suitable	100

Table 10. The groundwater samples classification in the Kasganj, Uttar Pradesh, North India for irrigation purposes^{27,25,2}.

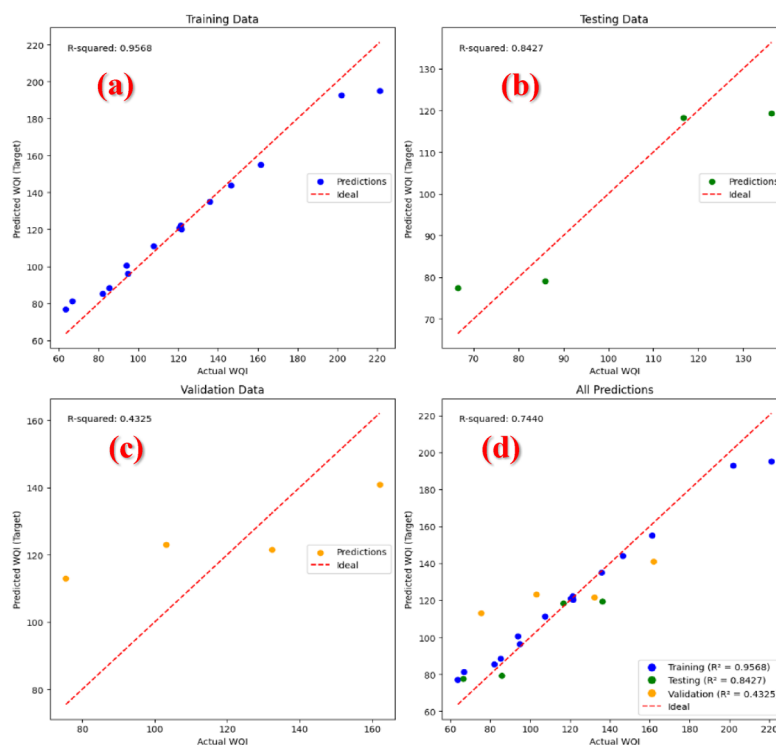


Fig. 11. (a–d) Regression of XGBoost model during training, testing and validation.

can be improved by integrating machine learning, offering a more reliable and effective means of water quality monitoring. The spatial distribution maps and statistical analysis clearly indicate significant hydrogeochemical heterogeneity within the study area, with certain parameters like fluoride, chloride, and sulfate showing concentrated zones of high values.

When considered in the context of recent related works, the results of this study are further supported. In Kerala, Aju et al.⁷² used machine learning models to predict groundwater quality and discovered that RF was the most successful, with an R^2 of 0.922⁷². Similarly, for groundwater forecasting, Hussein et al.⁷³ emphasized the superior predictive stability of RF and XGBoost over traditional models. The usefulness of ANN-based hybrid approaches for probabilistic risk assessment in fluoride-endemic areas was further illustrated by Islam et al.³¹. By combining WQI-based evaluation with spatial distribution analysis, the proposed work not only validates the effectiveness of RF in managing hydrogeochemical heterogeneity but also advances the field in comparison to these studies. This thorough approach highlights our methodology's contribution to sustainable groundwater monitoring and emphasizes its dependability and practical applicability.

Conclusion

The groundwater quality has been deteriorating from the geogenic and anthropogenic sources. However, the 115 water samples of twenty-three different locations of Kasganj reveal that the study region is under a serious

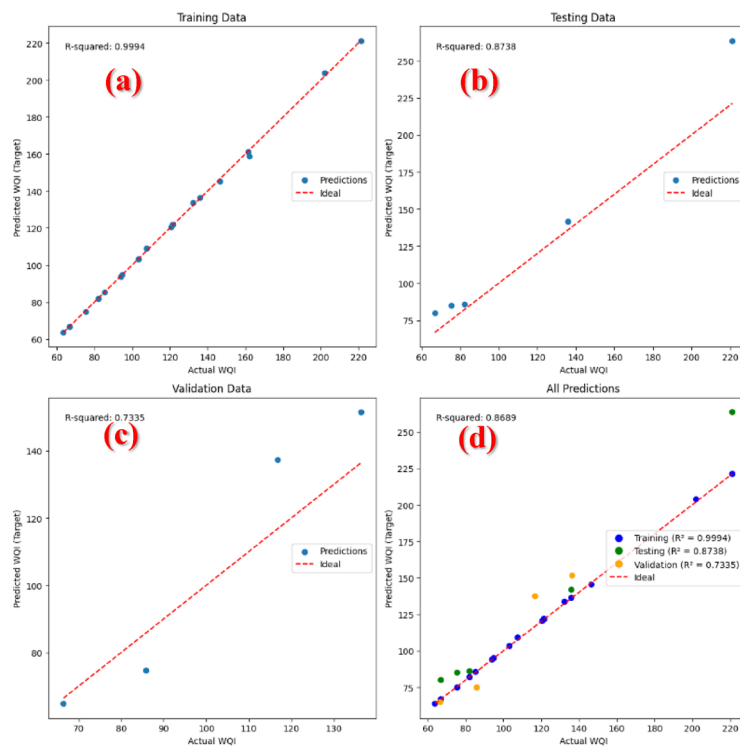


Fig. 12. (a–d) Regression of ANN model during training, testing and validation.

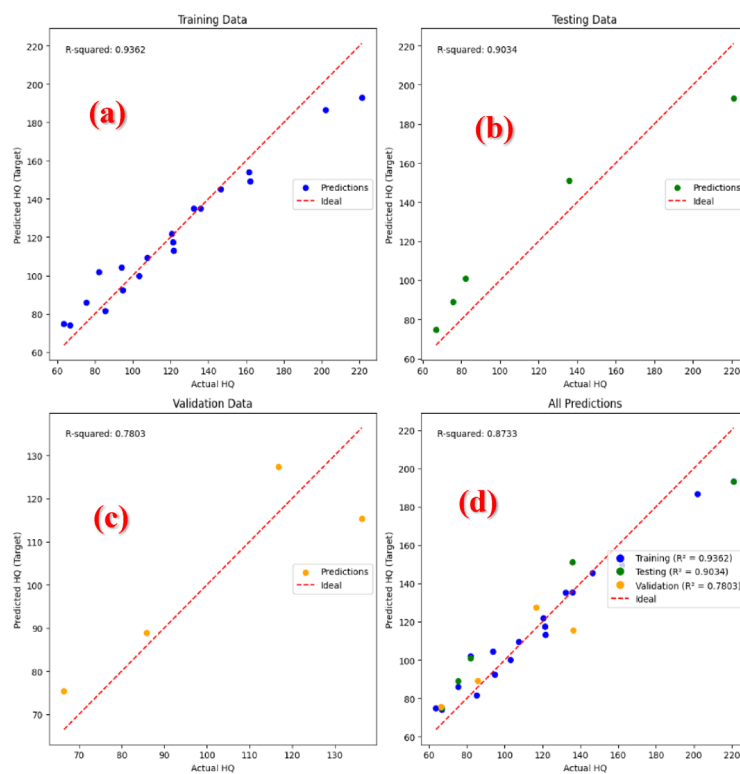


Fig. 13. (a–d) Regression of RF model during training, testing and validation.

Metrics	Model	Version				
		1	2	3	4	5
RSME	XGB	34.15	10.04	16.37	15.00	12.08
	ANN	11.88	22.15	18.88	68.03	13.95
	RF	30.83	6.38	25.18	16.92	5.97
MSE	XGB	1166.39	100.81	267.88	225.01	146.02
	ANN	141.11	490.54	356.41	4627.86	53.08
	RF	950.42	40.76	634.22	286.16	35.69
MAE	XGB	24.23	8.77	14.27	14.35	8.95
	ANN	11.19	19.72	14.14	53.08	12.12
	RF	24.53	3.77	19.70	16.58	5.49
RSQU	XGB	0.65	0.76	0.81	0.83	0.80
	ANN	0.9577	−0.18	0.75	−2.45	0.73
	RF	0.72	0.90	0.55	0.79	0.95

Table 11. Performance of XGB, ANN, and RF on metrics (RMSE, MSE, and MAE) for WQI.

threat to groundwater. Water Quality Indexing (WQI) and Irrigation Water Quality Indexing (IWQI) have also been utilized to distinguish the suitability and quality of water sites in the study area for affordable drinking and agricultural purposes. It has also been noticed that Total Dissolved Solids (TDS) and fluoride (F[−]) concentrations exceed WHO guidelines, posing significant health risks. Although pH and hardness were above permissible limits, which indicates the consistently elevated fluoride levels, in correlation with pH, alkalinity, and ion interactions (notably with hydrogen, sodium, and bicarbonate) and the geochemical mechanisms influencing groundwater chemistry in the region. Notably, 60.87% of the samples were classified as unsuitable for human consumption, with several falling into the “extremely poor” category. This highlights both the health risks and the urgent need for sustainable groundwater management. The predictive models used for assessing the water quality include Random Forest (RF), Artificial Neural Network (ANN), and XGBoost (XGB), and affirm that the RF model demonstrated the most balanced and reliable performance, achieving the lowest error metrics (RMSE: 5.97, MSE: 35.69, MAE: 5.49) and a strong coefficient of determination ($R^2=0.951$). While ANN slightly outperformed RF in R^2 (0.957), its higher error rates rendered RF the more robust choice overall. These machine learning models highlight the strong potential of accurately predicting and monitoring groundwater quality and offering valuable support for water resource management and public health strategies. The wide variation in water quality across the study area suggests it is influenced by both natural geological conditions and human activities. Additionally, the negative saturation index values for minerals like fluoride indicate undersaturation, which may increase fluoride mobility and contribute to its elevated levels in groundwater. Overall, this research reveals serious groundwater quality issues in the study region, which demonstrates how data-driven approaches, especially machine learning, can offer practical solutions for better groundwater monitoring. However, the effectiveness of these models depends heavily on the quality and representativeness of the input data, and the complexity of some algorithms may pose challenges in terms of transparency and interpretability for stakeholders and decision-makers.

Data availability

The data will be provided on a request from the corresponding author.

Received: 18 May 2025; Accepted: 22 September 2025
Published online: 28 October 2025

References

1. Ali, S., Mohammadi, A. A., Ali, H., Alinejad, N. & Maroosi, M. Qualitative assessment of ground water using the water quality index from a part of Western Uttar Pradesh North India. *Desalin. Water Treat.* **252**, 332–338 (2022).
2. Morovati, R., Badeenezhad, A., Najafi, M. & Azhdarpoor, A. Investigating the correlation between chemical parameters, risk assessment, and sensitivity analysis of fluoride and nitrate in regional groundwater sources using Monte Carlo. *Environ. Geochem. Health* **46**, 5 (2024).
3. Shukla, S. & Saxena, A. Appraisal of groundwater quality with human health risk assessment in parts of Indo-Gangetic alluvial plain North India. *Arch. Environ. Contam. Toxicol.* **80**, 55–73 (2021).
4. Li, P., Tian, R., Xue, C. & Wu, J. Progress, opportunities, and key fields for groundwater quality research under the impacts of human activities in China with a special focus on Western China. *Environ. Sci. Pollut. Res.* **24**, 13224–13234 (2017).
5. Gani, A., Singh, M., Pathak, S. & Hussain, A. Groundwater quality index development using the ANN model of Delhi metropolitan city India. *Environ. Sci. Pollut. Res.* <https://doi.org/10.1007/s11356-023-31584-4> (2023).
6. Ali, S. et al. Groundwater quality assessment using water quality index and principal component analysis in the Achnera block, Agra district, Uttar Pradesh Northern India. *Sci. Rep.* **14**, 1–13 (2024).
7. Panneerselvam, B., Muniraj, K., Pande, C. & Ravichandran, N. Prediction and evaluation of groundwater characteristics using the radial basic model in Semi-arid region India. *Int. J. Environ. Anal. Chem.* **103**, 1377–1393 (2023).
8. Sankar, K. et al. Integrated hydrogeophysical and GIS based demarcation of groundwater potential and vulnerability zones in a hard rock and sedimentary terrain of Southern India. *Chemosphere* **316**, 137305 (2023).
9. Adimalla, N., Li, P. & Venkatayogi, S. Hydrogeochemical evaluation of groundwater quality for drinking and irrigation purposes and integrated interpretation with water quality index studies. *Environ. Process.* **5**, 363–383 (2018).

10. Singh, S. & Hussian, A. Water quality index development for groundwater quality assessment of Greater Noida sub-basin, Uttar Pradesh India. *Cogent Eng.* **3**, 1177155 (2016).
11. Salehi, S., Chizari, M., Sadighi, H. & Bijani, M. Assessment of agricultural groundwater users in Iran: A cultural environmental bias. *Hydrogeol. J.* **26**, 285–295 (2018).
12. Takdastan, A. et al. Neuro-fuzzy inference system prediction of stability indices and sodium absorption ratio in Lordegan rural drinking water resources in west Iran. *Data Br.* **18**, 255–261 (2018).
13. Shams, M., Mohamadi, A. & Sajadi, S. A. Evaluation of corrosion and scaling potential of water in rural water supply distribution networks of Tabas Iran. *World Appl. Sci. J.* **17**, 1484–1489 (2012).
14. Ali, S., Gupta, S. K., Sinha, A., Khan, S. U. & Ali, H. Health risk assessment due to fluoride contamination in groundwater of Bichpuri, Agra, India: A case study. *Model. Earth Syst. Environ.* **8**, 299–307 (2022).
15. Faraji, H. et al. Correlation between fluoride in drinking water and its levels in breast milk in Golestan province Northern Iran. *Iran. J. Public Health* **43**, 1664–1668 (2014).
16. Berman, J. WHO: Waterborne disease is world's leading killer. **29**, 12 (2009).
17. Malik, A., Yasar, A., Tabinda, A. B. & Abubakar, M. Water-borne diseases, cost of illness and willingness to pay for diseases interventions in rural communities of developing countries. *Iran. J. Public Health* **41**, 39–49 (2012).
18. Khan, S. U., Asif, M., Alam, F., Khan, N. A. & Farooqi, I. H. Optimizing fluoride removal and energy consumption in a batch reactor using electrocoagulation: A smart treatment technology. 767–778 (2020). https://doi.org/10.1007/978-981-15-2545-2_62.
19. Singh, P. K., Tiwari, A. K. & Mahato, M. K. Qualitative assessment of surface water of West Bokaro coalfield, Jharkhand by using water quality index method. *Int. J. ChemTech Res.* **5**, 2351–2356 (2013).
20. Zhang, Q., Qian, H., Xu, P., Hou, K. & Yang, F. Groundwater quality assessment using a new integrated-weight water quality index (IWQI) and driver analysis in the Jiaokou irrigation district China. *Ecotoxicol. Environ. Saf.* **212**, 111992 (2021).
21. Dash, S. & Kalamdhad, A. S. Hydrochemical dynamics of water quality for irrigation use and introducing a new water quality index incorporating multivariate statistics. *Environ. Earth Sci.* **80**, 73 (2021).
22. Verma, P., Singh, P. K., Sinha, R. R. & Tiwari, A. K. Assessment of groundwater quality status by using water quality index (WQI) and geographic information system (GIS) approaches: A case study of the Bokaro district India. *Appl. Water Sci.* **10**, 27 (2020).
23. Chakraborty, B. et al. Geospatial assessment of groundwater quality for drinking through water quality index and human health risk index in an upland area of Chota Nagpur Plateau of West Bengal, India. 327–358 (2021). https://doi.org/10.1007/978-3-030-63422-3_19.
24. Ketata, M., Gueddari, M. & Bouhlila, R. Use of geographical information system and water quality index to assess groundwater quality in El Khairat deep aquifer (Enfidha, Central East Tunisia). *Arab. J. Geosci.* **5**, 1379–1390 (2012).
25. Omran, E.-S.E. A proposed model to assess and map irrigation water well suitability using geospatial analysis. *Water* **4**, 545–567 (2012).
26. Yang, Q., Li, Z., Xie, C., Liang, J. & Ma, H. Risk assessment of groundwater hydrochemistry for irrigation suitability in Ordos Basin China. *Nat. Hazards* **101**, 309–325 (2020).
27. Etteieb, S., Cherif, S. & Tarhouni, J. Hydrochemical assessment of water quality for irrigation: A case study of the Medjerda river in Tunisia. *Appl. Water Sci.* **7**, 469–480 (2017).
28. Brhane, G. K. Irrigation water quality index and GIS approach based groundwater quality assessment and evaluation for irrigation purpose in Ganta Afshum selected Kebeles, Northern Ethiopia. *Int. J. Emerg. Trends Sci. Technol.* <https://doi.org/10.18535/ijetst/v3i09.10> (2016).
29. Morovati, R., Abbasi, F., Samaei, M. R., Mehrazmay, H. & Lari, A. R. Modelling of n-Hexadecane bioremediation from soil by slurry bioreactors using artificial neural network method. *Sci. Rep.* **12**, 19662 (2022).
30. Hussein, E. A., Thron, C., Ghaziasgar, M., Bagula, A. & Vaccari, M. Groundwater prediction using machine-learning tools. *Algorithms* **13**, 300 (2020).
31. Islam, R. et al. Heliyon application of Monte Carlo simulation and artificial neural network model to probabilistic health risk assessment in fluoride-endemic areas. *Heliyon* **10**, e40887 (2024).
32. District Etah in Parts of the Central. **2**, 364–369 (2004).
33. Clesceri, L. S., Greenberg, A. E. & Trussell, R. R. *Standard Methods for the Examination of Water and Wastewater: Washington DC, American Public Health Association. Standard Methods for the Examination of Water and Wastewater, American Public Health Association* (1990).
34. Ali, S. et al. Variability of groundwater fluoride and its proportionate risk quantification via Monte Carlo simulation in rural and urban areas of Agra district India. *Sci. Rep.* **13**, 1–13 (2023).
35. Parkhurst, D. L. & Appelo, C. A. J. User's guide to PHREEQC (version 2): A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. *Water. Resour. Investig. Rep.* <https://doi.org/10.3133/wri994259> (1999).
36. Ako, A. A. et al. Spring water quality and usability in the Mount Cameroon area revealed by hydrogeochemistry. *Environ. Geochem. Health* **34**, 615–639 (2012).
37. WHO. Guidelines for Drinking-water Quality: Second addendum. *World Heal. Organ. Press* **1**, 17–19 (2008).
38. BIS. Indian Standard Drinking Water Specification (Second Revision). *Bur. Indian Stand.* **IS 10500**, 1–11 (2012).
39. Abbasnia, A. et al. Evaluation of groundwater quality using water quality index and its suitability for assessing water for drinking and irrigation purposes: Case study of Sistan and Baluchistan province (Iran). *Hum. Ecol. Risk Assess. Int. J.* **25**, 988–1005 (2019).
40. Balan, In., Madan Kumar, P. & Shivakumar, M. An assessment of groundwater quality using water quality index in Chennai, Tamil Nadu India. *Chron. Young Sci.* **3**, 146 (2012).
41. Brown, R. M., McClelland, N. I., Deininger, R. A. & O'Connor, M. F. A water quality index-crashing the psychological barrier. *Adv. Water Pollut. Res.* <https://doi.org/10.1016/b978-0-08-017005-3.50067-0> (1973).
42. Farmaki, E. G., Thomaidis, N. S. & Efsthathiou, C. E. Artificial neural networks in water analysis: Theory and applications. *Int. J. Environ. Anal. Chem.* **90**, 85–105 (2010).
43. Uddin, M. G. et al. Assessment of human health risk from potentially toxic elements and predicting groundwater contamination using machine learning approaches. *J. Contam. Hydrol.* **261**, 104307 (2024).
44. Zhu, M. et al. A review of the application of machine learning in water quality evaluation. *Eco-Environment Heal.* **1**, 107–116 (2022).
45. Li, J. et al. Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmos. Res.* **276**, 106238 (2022).
46. Wang, D. et al. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* **301**, 113941 (2022).
47. Nong, X., Shao, D., Zhong, H. & Liang, J. Evaluation of water quality in the South-to-North water diversion project of China using the water quality index (WQI) method. *Water Res.* **178**, 115781 (2020).
48. Chen, T. & Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). <https://doi.org/10.1145/2939672.2939785>.
49. Niazkar, M. et al. Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environ. Model. Softw.* **174**, 105971 (2024).
50. Wang, X., Tian, Y. & Liu, C. Assessment of groundwater quality in a highly urbanized coastal city using water quality index model and bayesian model averaging. *Front. Environ. Sci.* **11**, 1–11 (2023).

51. Osman, A. I., Najah Ahmed, A., Chow, M. F., Feng Huang, Y. & El-Shafie, A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Eng. J.* **12**, 1545–1556 (2021).
52. Abernethy, J. et al. Flint Water Crisis: Data-Driven Risk Assessment Via Residential Water Testing. (2016).
53. Kuo, J.-T., Hsieh, M.-H., Lung, W.-S. & She, N. Using artificial neural network for reservoir eutrophication prediction. *Ecol. Modell.* **200**, 171–177 (2007).
54. Ghiasi, B. et al. Uncertainty quantification of granular computing-neural network model for prediction of pollutant longitudinal dispersion coefficient in aquatic streams. *Sci. Rep.* **12**, 4610 (2022).
55. Mohammadi, A. A., Ghaderpoori, M., Yousefi, M., Rahmatipoor, M. & Javan, S. Prediction and modeling of fluoride concentrations in groundwater resources using an artificial neural network: A case study in Khaf. *Environ. Heal. Eng. Manag.* **3**, 217–224 (2016).
56. Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H. & Ramli, M. F. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Mar. Pollut. Bull.* **64**, 2409–2420 (2012).
57. Al-Adhaileh, M. H. & Alsaade, F. W. Kullanarak Su Kalitesinin Modellenmesi ve Tahminiyapay zeka—Modelling and prediction of water quality by using artificial intelligence. *Sustain.* **13**, 1–18 (2021).
58. Uddin, M. G., Nash, S., Mahammad Diganta, M. T., Rahman, A. & Olbert, A. I. Robust machine learning algorithms for predicting coastal water quality index. *J. Environ. Manag.* **321**, 115923 (2022).
59. Singh, B., Sihag, P., Singh, V. P., Sepahvand, A. & Singh, K. Soft computing technique-based prediction of water quality index. *Water Supply* **21**, 4015–4029 (2021).
60. Voyant, C. et al. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **105**, 569–582 (2017).
61. Farnaaz, N. & Jabbar, M. A. Random forest modeling for network intrusion detection system. *Procedia Comput. Sci.* **89**, 213–217 (2016).
62. Zhang, K. et al. Identification of anthropogenic and natural inputs of sulfate into river system of carbonate Zn-Pb mining area in Southwest China: Evidence from hydrochemical composition, $\delta^{34}\text{S}\text{SO}_4$ and $\delta^{18}\text{O}\text{SO}_4$. *Water* **16**, 2311 (2024).
63. Rao, S., Mogili, N. V., Priscilla, A. & Lydia, A. Aqueous chemistry of anthropogenically contaminated Bengaluru lakes. *Sustain. Environ. Res.* **30**, 8 (2020).
64. Sharma, M. K. & Kumar, M. Sulphate contamination in groundwater and its remediation: An overview. *Environ. Monit. Assess.* **192**, 74 (2020).
65. Zhou, L. et al. Hydrogeochemistry of fluoride in shallow groundwater of the abandoned Yellow river delta China. *Hydrol. Res.* **52**, 572–584 (2021).
66. Kumar, P. S. Fluoride in groundwater—Sources, geochemical mobilization and treatment options. *Int. J. Environ. Sci. Nat. Resour.* **1**, 106–108 (2017).
67. Ahmed, S., Khurshid, S., Sultan, W. & Shadab, M. B. Statistical analysis and water quality index development using GIS of Mathura city, Uttar Pradesh, India. *Desalin. Water Treat.* **177**, 152–166 (2020).
68. Morovati, R., Badeenezhad, A., Najafi, M. & Azhdarpoor, A. Investigating the correlation between chemical parameters, risk assessment, and sensitivity analysis of fluoride and nitrate in regional groundwater sources using Monte Carlo. *Environ. Geochem. Health* **46**, 1–17 (2024).
69. Kumar, A., Maraju, S. & Bhat, A. Application of ArcGIS geostatistical analyst for interpolating environmental data from observations. *Environ. Prog.* **26**, 220–225 (2007).
70. Taloor, A. K. et al. Spring water quality and discharge assessment in the Basantar watershed of Jammu Himalaya using geographic information system (GIS) and water quality Index(WQI). *Groundw. Sustain. Dev.* **10**, 100364 (2020).
71. Balamurugan, P., Kumar, P. S., Shankar, K., Nagavinothini, R. & Vijayasurya, K. Non-carcinogenic risk assessment of groundwater in Southern part of Salem district in Tamilnadu India. *J. Chil. Chem. Soc.* **65**, 4697–4707 (2020).
72. Aju, C. D. et al. Groundwater quality prediction and risk assessment in Kerala, India: A machine-learning approach. *J. Environ. Manag.* **370**, 122616 (2024).
73. Hussein, E. A., Thron, C., Ghaziasgar, M., Bagula, A. & Vaccari, M. Groundwater prediction using machine-learning tools. *Algorithms* **13**, 1–16 (2020).

Acknowledgements

The researchers would like to acknowledge Indian Institute of Technology (Indian School of Mines), Dhanbad and Sharda University Agra for running the testing facilities used throughout the study.

Author contributions

Raisul Islam, Alok Sinha and Shahjad Ali conducted the comparative analysis of machine learning models, designed the experimental framework, and conducted the experiments. Kamlesh Deshmukh, and Salman Ahmed worked together in a collaborative spirit to handle the data analysis and interpretation. Athar Hussain contributed essential reagents and materials required for the study. Rajesh Kumar Deolia, Mohammad Usama and Jitendra Kumar evaluated the performance of lab works and mathematical calculations.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.A.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025