International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

# Integrated Groundwater Quality Assessment and Machine Learning Prediction in Central Uttar Pradesh, India

**Nitin Mishra**
Department of Civil Engineering, Institute of Engineering and Technology, Lucknow-226021, India

**Abstract- Groundwater is the principal source of drinking and irrigation water in the Indo-Gangetic alluvial plains of Uttar Pradesh, India. Rapid urbanization, agricultural intensification, excessive groundwater abstraction, and geogenic contamination have significantly affected groundwater quality in the region. The present study evaluates groundwater quality in Central Uttar Pradesh using hydrogeochemical assessment, entropy-weighted water quality index (EWQI), and machine learning (ML) prediction techniques. A total of 178 groundwater samples were analyzed for major physicochemical parameters including pH, EC, TDS, TH, Ca2+, Mg2+, Na+, K+, HCO3−, Cl−, SO42−, NO3−, F−, SiO2, and CO32−. The entropy weight method was employed to minimize subjectivity in water quality assessment, while hydrogeochemical interpretations were carried out using Piper and Gibbs diagrams. Three machine learning models, namely Classification and Regression Tree (CART), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), were implemented to predict groundwater quality conditions. The results revealed that groundwater chemistry is predominantly controlled by rock–water interaction and ion exchange processes, with Ca–HCO3 and mixed hydrochemical facies dominating the study area. The EWQI values indicated that most groundwater samples fall within good to medium drinking water quality categories, although localized fluoride enrichment was observed in several locations. Among the applied models, XGBoost demonstrated superior predictive capability with R2 = 0.9597, RMSE = 2.2376, and MAE = 1.7690, outperforming RF and CART models. The findings highlight the effectiveness of integrating GIS-based hydrogeochemical analysis with machine learning approaches for groundwater quality prediction and sustainable groundwater management in Central Uttar Pradesh.**

**Keywords- Groundwater quality, EWQI, Machine learning, XGBoost, Random Forest, Hydrogeochemistry, Fluoride contamination, Central Uttar Pradesh.**

## I. INTRODUCTION

Groundwater is one of the most important freshwater resources globally and plays a critical role in sustaining domestic, agricultural, and industrial activities. In densely populated regions such as the Indo-Gangetic plains of India, groundwater has become the primary source of potable water because of increasing pressure on surface water resources. However, groundwater systems are increasingly threatened by anthropogenic activities, population growth, agricultural intensification, and uncontrolled extraction (Gleeson et al., 2012; Rodell et al., 2009). In India, especially in Uttar Pradesh, groundwater quality deterioration has emerged as a major environmental and public health concern.

The alluvial aquifers of Uttar Pradesh are characterized by complex hydrogeochemical interactions influenced by lithology, agricultural runoff, urban wastewater discharge, and geogenic processes. Several studies have reported elevated concentrations of fluoride, nitrate, salinity, and hardness in groundwater across the Indo-Gangetic basin (Adimalla, 2020; Rahman et al., 2021). Long-term exposure to contaminated groundwater may lead to severe health risks such as fluorosis, kidney disorders, and gastrointestinal diseases.

Water quality indices (WQIs) have been widely used to simplify complex hydrochemical information into a single numerical value for evaluating groundwater suitability for drinking purposes (Brown et al., 1970; Horton, 1965). Traditional WQI approaches often involve subjective

assignment of parameter weights, which may introduce uncertainties in assessment. To overcome this limitation, entropy-based weighting methods have been increasingly adopted because they provide objective parameter weighting based on information entropy theory (Li et al., 2019).

In recent years, machine learning (ML) techniques have gained significant attention for environmental prediction and groundwater quality modeling due to their ability to capture nonlinear relationships among hydrochemical variables (Adhikari & Hartemink, 2016; Arabameri et al., 2020). Algorithms such as Random Forest (RF), Classification and Regression Tree (CART), and Extreme Gradient Boosting (XGBoost) have demonstrated excellent predictive performance in groundwater quality assessment studies (Singha et al., 2021; Yadav et al., 2024; Karimi et al., 2025).

Several recent studies have highlighted the integration of GIS and ML techniques for groundwater quality prediction. Abu El-Magd et al. (2023) developed integrated machine learning models for groundwater quality assessment using WQI approaches, while Singh et al. (2025) applied GIS-based ML models for fluoride prediction in groundwater systems. Similarly, Raheja et al. (2024) demonstrated the effectiveness of ensemble ML techniques in evaluating drinking water quality.

Despite increasing research efforts, limited studies have integrated entropy-weighted groundwater quality assessment with comparative machine learning prediction in Central Uttar

Pradesh. Therefore, the present study aims to:

1. Evaluate groundwater quality using hydrochemical and entropy-weighted water quality index approaches.
2. Identify dominant hydrogeochemical processes controlling groundwater chemistry.
3. Develop machine learning models (CART, RF, and XGBoost) for groundwater quality prediction.
4. Compare the predictive performance of the applied machine learning models.
5. Provide scientific insights for sustainable groundwater resource management in Central Uttar Pradesh.

The present research contributes to the growing body of literature on groundwater quality modeling by integrating hydrogeochemical interpretation with advanced machine learning techniques for regional-scale groundwater assessment.

## II. STUDY AREA

The study area is located in Central Uttar Pradesh, India, within the Indo-Gangetic alluvial plains. The region is characterized by flat topography, fertile alluvial deposits, and intensive agricultural activities. Groundwater serves as the major source of drinking and irrigation water for the local population.

The climate of the study area is subtropical with distinct summer, monsoon, and winter seasons. Average annual rainfall is mainly received during the southwest monsoon season. The geology is dominated by Quaternary alluvial deposits consisting of sand, silt, clay, and gravel layers. These unconsolidated sediments form highly productive aquifers with varying hydrochemical characteristics.

The hydrogeological conditions are influenced by recharge from rainfall, canal seepage, irrigation return flow, and river–groundwater interaction. Intensive agricultural practices involving fertilizers and pesticides may significantly influence groundwater chemistry.


Fig1. Study area map of Central Uttar Pradesh, India

## III. MATERIALS AND METHODS

**Groundwater Sampling and Data Collection**

A total of 178 groundwater samples were collected from different locations across Central Uttar Pradesh for hydrochemical analysis. The dataset included geographical coordinates and physicochemical parameters are pH, EC, TDS, TH, $Ca^{2+}$, $Mg^{2+}$, $Na^{+}$, $K^{+}$, $HCO_3^{-}$, $Cl^{-}$, $SO_4^{2-}$, $NO_3^{-}$, $F^{-}$, $SiO_2$, and $CO_3^{2-}$.The collected samples were analyzed using

standard analytical procedures recommended for groundwater quality assessment.

**Data Preprocessing**

The groundwater dataset was preprocessed prior to modeling and analysis. Missing values, below detection limit values, and inconsistent observations were handled carefully to ensure data quality. Numerical conversion and normalization techniques were applied to improve model performance.

Outlier removal and MinMax normalization were performed to standardize the dataset before machine learning implementation.

**Entropy Weighted Water Quality Index (EWQI)**

The Entropy Weighted Water Quality Index (EWQI) was employed to evaluate groundwater suitability for drinking purposes. The entropy method provides objective parameter weighting based on Shannon's entropy theory (Shannon, 1948).

**The entropy calculation procedure involved:**

1. Data normalization
2. Entropy calculation
3. Entropy weight determination
4. Quality rating calculation
5. EWQI estimation

The EWQI method reduces subjectivity associated with conventional WQI approaches and provides a more reliable assessment framework.

**Hydrogeochemical Analysis**

Hydrogeochemical characterization of groundwater was performed using Piper and Gibbs diagrams.

**Piper Diagram**

The Piper diagram was used to classify groundwater hydrochemical facies and identify dominant ionic compositions. The diagram assists in understanding geochemical evolution and mixing processes within aquifers.

**Machine Learning Models**

Three machine learning algorithms were implemented to predict groundwater quality conditions.

**Classification and Regression Tree (CART)**

CART is a tree-based machine learning algorithm that partitions the dataset into homogeneous subsets using recursive binary splitting (Breiman et al., 1984). CART models are computationally efficient and interpretable.

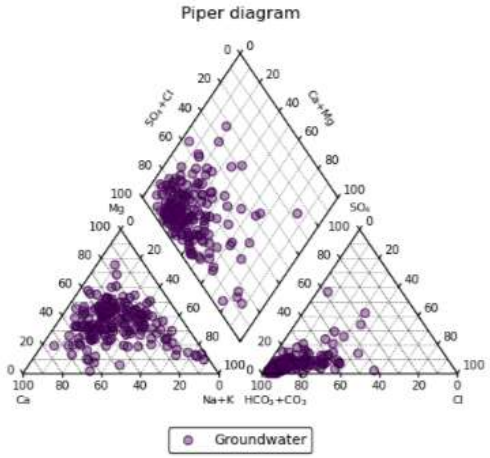



Fig 2. Piper Diagram showing classification of groundwater hydrochemical facies.

**Random Forest (RF)**

Random Forest is an ensemble learning technique based on multiple decision trees and bootstrap aggregation (Breiman, 2001). RF improves prediction accuracy and reduces overfitting.

**Extreme Gradient Boosting (XGBoost)**

XGBoost is an advanced gradient boosting framework designed for high-performance predictive modeling (Chen & Guestrin, 2016). The algorithm sequentially improves weak learners to minimize prediction errors.

**Model Evaluation**

The predictive performance of the models was evaluated using:

- Coefficient of Determination (R2)
- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)

Higher R2 and lower RMSE/MAE values indicate better model performance.

## IV. RESULTS AND DISCUSSION

**Groundwater Hydrochemistry**

The groundwater samples exhibited noticeable spatial variation in hydrochemical characteristics across the study area. The pH values indicated slightly alkaline groundwater conditions,

which are common in alluvial aquifers due to carbonate mineral dissolution.

Electrical conductivity and total dissolved solids showed moderate variability, reflecting differences in mineral dissolution, anthropogenic influence, and groundwater residence time. Elevated hardness values in several samples indicate significant contributions from calcium and magnesium-bearing minerals.

Fluoride concentrations varied spatially, with certain locations showing elevated levels that may pose long-term health risks if consumed continuously.

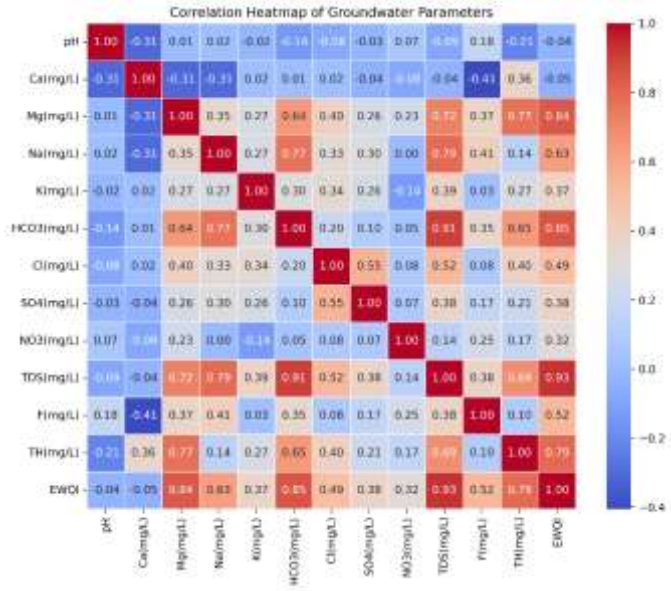



Fig 3. Correlation Heat map of groundwater parameters.

**Hydrochemical Facies**

The Piper diagram revealed that Ca–HCO3 and mixed hydrochemical facies dominate the study area. The predominance of bicarbonate-rich water suggests active carbonate weathering and rock–water interaction processes.

Mixed facies observed in several samples indicate ion exchange and transitional groundwater chemistry influenced by anthropogenic activities and aquifer heterogeneity.

**Gibbs Diagram Interpretation**

The Gibbs diagram indicated that rock–water interaction is the dominant mechanism controlling groundwater chemistry in the study area. Most samples clustered within the rock dominance zone, suggesting mineral weathering and dissolution processes as primary hydrochemical controls.

Limited samples showed influence of evaporation processes, particularly in areas with higher salinity and dissolved ion concentrations.

**Entropy Weighted Water Quality Index (EWQI)**

The EWQI analysis categorized groundwater quality into different drinking suitability classes. Most groundwater samples belonged to good and medium quality categories, indicating general suitability for drinking purposes after conventional treatment.

However, localized areas exhibiting elevated fluoride and dissolved ion concentrations require special attention and continuous monitoring.

The entropy weighting approach effectively minimized subjectivity in parameter weighting and provided a reliable framework for groundwater quality assessment.

**Machine Learning Prediction Performance**

The three machine learning models demonstrated strong predictive capability for groundwater quality estimation.

**CART Model Performance**

The CART model achieved satisfactory prediction accuracy with:

- R2 = 0.9327
- RMSE = 2.8943
- MAE = 2.4710

The model effectively captured nonlinear relationships between hydrochemical variables but exhibited slightly lower accuracy compared to ensemble-based methods.

**Random Forest Model Performance**

The Random Forest model demonstrated improved predictive performance:

- R2 = 0.9354
- RMSE = 2.8354
- MAE = 2.3688

The ensemble learning structure enhanced generalization capability and reduced model variance

.

**XGBoost Model Performance**

Among all models, XGBoost achieved the best predictive performance

- R2 = 0.9598
- RMSE = 2.2376
- MAE = 1.7690

The superior performance of XGBoost may be attributed to its gradient boosting framework, regularization capability, and efficient handling of nonlinear interactions.

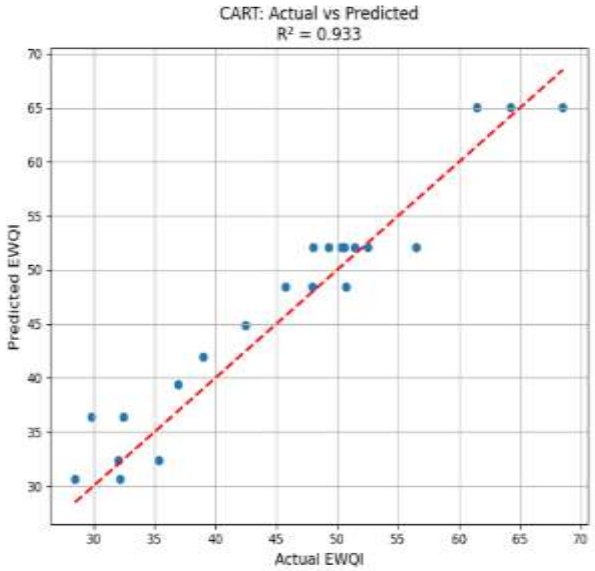
**Comparative Analysis of ML Models**

The comparative analysis clearly demonstrated the superiority of XGBoost over RF and CART models for groundwater quality prediction.

| Model | R2 | RMSE | MAE |
|---|---|---|---|
| CART | 0.9327 | 2.8943 | 2.4710 |
| Random Forest | 0.9354 | 2.8354 | 2.3688 |
| XGBoost | 0.9598 | 2.2376 | 1.7690 |

The high R2 values obtained for all models indicate strong predictive relationships between hydrochemical variables and groundwater quality conditions.

The findings are consistent with previous studies that reported superior performance of ensemble and boosting algorithms in groundwater quality prediction (Yadav et al., 2024; Karimi et al., 2025; Singha et al., 2021).



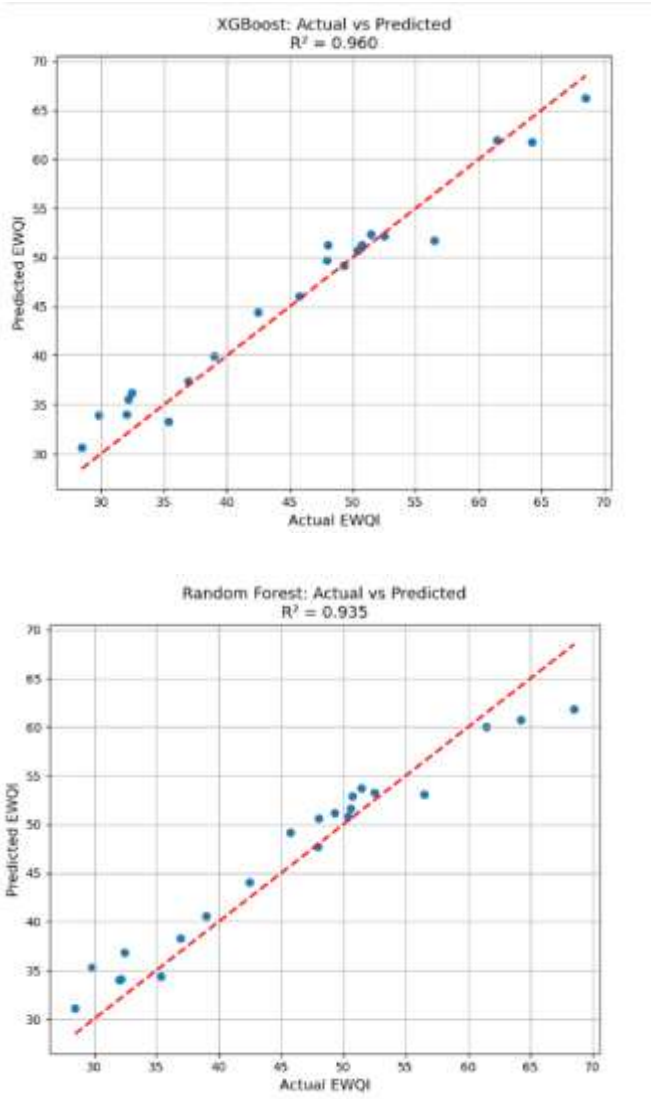




Fig.3. Plot showing prediction versus test data of (a)CART, (b)RF, (c)XGBoost

**Implications for Groundwater Management**

The integration of hydrogeochemical assessment and machine learning modeling provides an efficient framework for sustainable groundwater management.

**The developed ML models can support:**

- Rapid groundwater quality prediction
- Identification of vulnerable zones
- Decision-making for groundwater monitoring
- Sustainable drinking water planning
- Long-term groundwater resource management

The methodology developed in this study can be applied to other alluvial aquifer systems facing groundwater quality challenges.

## V. CONCLUSION

The present study integrated hydrogeochemical analysis, entropy-weighted water quality assessment, and machine learning modeling to evaluate groundwater quality in Central Uttar Pradesh, India.

The hydrochemical analysis revealed that groundwater chemistry is predominantly governed by rock–water interaction processes, with Ca–HCO3 and mixed facies dominating the aquifer system. The EWQI assessment indicated that most groundwater samples fall within good to medium drinking water quality categories, although localized fluoride enrichment was observed in certain areas.

Three machine learning models, namely CART, Random Forest, and XGBoost, were successfully implemented for groundwater quality prediction. Among the applied algorithms, XGBoost exhibited the best predictive performance with R2 = 0.9598, RMSE = 2.2376, and MAE = 1.7690.

The study demonstrates that integrating hydrogeochemical techniques with advanced machine learning models can significantly improve groundwater quality prediction and assessment. The developed framework can assist policymakers, environmental agencies, and water resource managers in sustainable groundwater planning and contamination mitigation.

Future studies may incorporate deep learning approaches, temporal groundwater datasets, remote sensing variables, and spatial interpolation techniques for further improvement in groundwater quality prediction accuracy.

**Acknowledgements**
The authors acknowledge the support of groundwater data collection agency (CGWB) and laboratory facilities used for hydrochemical analysis. The authors are also grateful to the open-source scientific computing community for providing computational tools used in this research.

**Conflict of Interest**
The authors declare no conflict of interest.

**Data Availability Statement**
The data used in this study are available from CGWB groundwater manual (2023-24), Uttar Pradesh.

## REFERENCES

1. Abu El-Magd, S. A., Ismael, I. S., El-Sabri, M. A. S., Abdo, M. S., & Farhat, H. I. (2023). Integrated machine learning–based model and WQI for groundwater quality assessment. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-023-25938-1
2. Adhikari, K., & Hartemink, A. E. (2016). Machine learning applications in environmental sciences. Geoderma, 265, 17–26. https://doi.org/10.1016/j.geoderma.2015.11.009
3. Adimalla, N. (2020). Groundwater quality assessment using WQI and GIS techniques. Environmental Earth Sciences, 79, 1–15. https://doi.org/10.1007/s12665-020-09089-9
4. Adimalla, N., & Wu, J. (2019). Groundwater quality assessment using GIS techniques. Environmental Monitoring and Assessment, 191, 1–17. https://doi.org/10.1007/s10661-019-7313-8
5. Arabameri, A., et al. (2020). Modeling groundwater quality using machine learning techniques. Science of the Total Environment, 703, 135593. https://doi.org/10.1016/j.scitotenv.2019.135593
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
7. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. https://doi.org/10.1201/9781315139470
8. Brown, R. M., McClelland, N. I., Deininger, R. A., & O'Connor, M. F. (1970). A water quality index—Do we dare? Water and Sewage Works, 117(10), 339–343.
9. Busico, G., et al. (2020). Hydrogeochemical processes influencing groundwater quality. Science of the Total Environment, 713, 136718. https://doi.org/10.1016/j.scitotenv.2020.136718
10. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference. https://doi.org/10.1145/2939672.2939785
11. Das, C. R., & Das, S. (2026). Groundwater quality assessment for drinking by weighted WQIs: A comprehensive review. Environmental Earth Sciences, 85, 90. https://doi.org/10.1007/s12665-026-12823-6

12. Egbueri, J. C., & Agbasi, J. C. (2022). Combining data-intelligent algorithms for assessment and predictive modeling of groundwater quality. Environmental Science and Pollution Research, 29, 57147–57171. https://doi.org/10.1007/s11356-022-19818-3
13. Gleeson, T., et al. (2012). Groundwater sustainability. Nature, 488, 197–200. https://doi.org/10.1038/nature11295
14. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. https://doi.org/10.1007/978-0-387-84858-7
15. Horton, R. K. (1965). An index number system for rating water quality. Journal of the Water Pollution Control Federation, 37(3), 300–306.
16. Karimi, H., Sahour, S., Khanbeyki, M., Gholami, V., Sahour, H., Shahabi-Ghahfarokhi, S., & Mohammadi, M. (2025). Enhancing groundwater quality prediction through ensemble machine learning techniques. Environmental Monitoring and Assessment, 197, 21. https://doi.org/10.1007/s10661-024-13506-0
17. Kaur, H., Bansod, B. S., Khungar, P., & Dhawan, C. (2025). Combining clustering and ensemble learning for groundwater quality monitoring: A data-driven framework. Environmental Science and Pollution Research, 32, 13862–13903. https://doi.org/10.1007/s11356-025-36477-2
18. Lakshmi, L. B., Rao, P. R., Mohan, C. C., Kumar, L. K., Kumar, K. S., & Kumar, B. V. S. (2024). Prediction of physico-chemical characteristics of groundwater using machine learning model. Advances in Computer Science Research. https://doi.org/10.2991/978-94-6463-471-6_60
19. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539
20. Li, P., Qian, H., & Wu, J. (2019). Application of entropy weight method in groundwater quality assessment. Environmental Science and Pollution Research, 26, 154–164. https://doi.org/10.1007/s11356-018-3656-3
21. MacDonald, A. M., et al. (2016). Groundwater resources globally. Nature Geoscience, 9, 217–224. https://doi.org/10.1038/ngeo2590
22. Rahman, M. M., et al. (2021). Groundwater contamination and human health risks. Environmental Research, 194, 110615. https://doi.org/10.1016/j.envres.2020.110615
23. Raheja, H., Goel, A., & Pal, M. (2024). Evaluation of groundwater quality for drinking purposes based on machine learning algorithms and GIS. Sustainable Water Resources Management, 10, 11. https://doi.org/10.1007/s40899-023-00990-4
24. Rodell, M., et al. (2009). Satellite-based groundwater depletion study. Nature, 460, 999–1002. https://doi.org/10.1038/nature08238
25. Sener, S., Sener, E., & Davraz, A. (2017). Evaluation of groundwater quality using WQI. Environmental Earth Sciences, 76, 1–13. https://doi.org/10.1007/s12665-017-6507-6
26. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
27. Shams, R., et al. (2024). Application of ML models for water quality prediction. Environmental Modelling & Software, 170, 105756. https://doi.org/10.1016/j.envsoft.2023.105756
28. Singh, R., Tripathy, S., Gupta, A. K., Uddameri, V., & Bagal, S. R. (2025). Enhancing spatial interpretability of fluoride in groundwater using an integrated GIS and machine learning approach. Earth Systems and Environment. https://doi.org/10.1007/s41748-025-00657-4
29. Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. Chemosphere, 276, 130265. https://doi.org/10.1016/j.chemosphere.2021.130265
30. Subramani, T., et al. (2005). Groundwater quality evaluation and hydrogeochemistry. Environmental Geology, 47, 109–119. https://doi.org/10.1007/s00254-004-1122-6
31. Torres-Martínez, J. A., et al. (2024). Machine learning approaches for groundwater quality prediction. Journal of Hydrology, 629, 130539. https://doi.org/10.1016/j.jhydrol.2023.130539
32. Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their application. Ecological Indicators, 122, 107218. https://doi.org/10.1016/j.ecolind.2020.107218
33. Yang, H., Jia, C., Yang, F., Yang, X., & Wei, R. (2023). Water quality assessment of deep learning–improved comprehensive pollution index. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-023-27174-z
34. Yadav, A., Raj, A., & Yadav, B. (2024). Enhancing local-scale groundwater quality predictions using advanced

machine learning approaches. Journal of Environmental Management, 370, 122903. https://doi.org/10.1016/j.jenvman.2024.122903