Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 159-167

Machine Learning–Based Water Quality Index for Spatio-Temporal Groundwater Quality Assessment in Rewari District, India

Manju Kumari[1], R.C. Chhipa[1], Anil Kumar Bansal[2]
[1]Department Chemistry, Suresh Gyan Vihar University, Jaipur, India
[2]Department Chemistry, Agrawal PG College, Jaipur, India

Email: makuyadav24@gmail.com, rc.chhipa@mygyanvihar.com, bansal.anilkumar@gmail.com

Abstract: Groundwater quality in industrial regions is increasingly threatened by anthropogenic activities, necessitating robust and data-driven assessment frameworks beyond conventional index-based approaches. This paper proposes a Machine Learning–Based Water Quality Index (ML-WQI) for comprehensive spatio-temporal assessment of groundwater quality in the industrial regions of Rewari District, Haryana, India. Physicochemical groundwater parameters, including pH, total dissolved solids, electrical conductivity, nitrate, fluoride, and chloride, were analyzed using supervised machine learning models to capture nonlinear relationships among water quality indicators. Multiple learning algorithms were evaluated, and the proposed ML-WQI framework demonstrated superior predictive performance compared to traditional WQI methods, achieving improved accuracy, reduced error metrics, and higher robustness. Spatial analysis identified distinct contamination hotspots within major industrial clusters, while temporal evaluation revealed a progressive deterioration in groundwater quality over recent years. The results confirm that the ML-WQI effectively captures both spatial heterogeneity and temporal variability in groundwater quality influenced by industrial activities. The proposed framework provides a scalable and reliable decision-support tool for groundwater monitoring, environmental management, and policy formulation in industrially impacted regions.

Keywords: Machine Learning, Water Quality Index, Groundwater Quality Assessment, Spatio-Temporal Analysis, Industrial Pollution

1. Introduction

Groundwater is a critical source of freshwater for domestic, agricultural, and industrial use, particularly in semi-arid regions of India where surface water resources are limited [1]. Rapid industrialization, urban expansion, and intensive agricultural practices have significantly altered the natural hydrogeochemical balance, leading to the deterioration of groundwater quality. Industrial regions are especially vulnerable due to the discharge of untreated or partially treated effluents, leaching of contaminants, and overexploitation of aquifers [2]. These challenges necessitate reliable and efficient approaches for continuous groundwater quality assessment and management.

Conventional groundwater quality assessment methods often rely on individual physicochemical parameters or traditional WQI models that use fixed weights and linear aggregation techniques [3]. While such approaches provide a simplified representation of water quality, they are limited in their ability to capture complex, nonlinear interactions among multiple water quality parameters and their spatial and temporal variability. Moreover, conventional WQI methods may not adequately reflect localized contamination patterns, particularly in industrially impacted regions where pollutant dynamics are highly heterogeneous.

Recent advances in ML have demonstrated significant potential in environmental monitoring and water resource management. ML models are capable of learning complex, nonlinear relationships from large datasets and have been successfully applied to water quality prediction, contamination source identification, and anomaly detection [4,5]. However, most existing studies focus primarily on parameter-wise prediction or short-term forecasting, with limited attention to the development of an integrated machine learning–based Water Quality Index that can systematically evaluate spatial and temporal groundwater quality variations.

Rewari District in Haryana, India, hosts several industrial clusters that have experienced rapid growth in recent years, raising concerns regarding groundwater contamination and long-term sustainability. Despite the growing environmental pressure, comprehensive spatio-temporal assessments of groundwater quality using advanced data-driven approaches remain limited for this region. A robust ML-driven WQI framework is therefore essential to accurately quantify groundwater quality degradation, identify contamination hotspots, and support evidence-based decision-making.

In this context, this paper proposes a ML-WQI for the spatio-temporal assessment of groundwater quality in the industrial regions of Rewari District, India. The proposed framework integrates multiple physicochemical groundwater parameters with supervised machine learning models to overcome the limitations of conventional WQI methods by effectively capturing nonlinear interactions among water quality indicators. The study contributes by developing an ML-driven WQI, performing spatial analysis to identify industrial contamination hotspots, and conducting temporal evaluation to assess long-term groundwater quality trends. Overall, the proposed ML-WQI framework provides a scalable and reliable tool for groundwater quality monitoring and supports sustainable groundwater management and evidence-based environmental policy formulation in industrially affected regions.

2. Literature Review

Groundwater quality assessment has traditionally been carried out using physicochemical analysis and index-based evaluation techniques to simplify complex water quality data into a single representative value. The WQ has been widely adopted due to its simplicity and interpretability, integrating multiple water quality parameters based on predefined weights and standards [6]. Several studies have employed conventional WQI models for groundwater assessment in industrial and urban regions, reporting their effectiveness in identifying general water quality status. However, these methods rely on linear aggregation and expert-assigned weights, which limit their ability to represent nonlinear interactions among water quality parameters and spatial heterogeneity in contaminated regions.

To overcome these limitations, statistical and multivariate techniques such as principal component analysis (PCA), factor analysis, and cluster analysis have been extensively applied to groundwater quality datasets [7[. These approaches assist in dimensionality reduction, source identification, and pattern recognition. While multivariate statistical methods provide valuable insights into groundwater pollution sources and variability, they are primarily exploratory and do not offer predictive capabilities or adaptive index computation for dynamic spatio-temporal assessment.

Recent advancements in ML have significantly enhanced groundwater quality modeling and prediction. Supervised learning algorithms such as artificial neural networks (ANN), support vector machines (SVM), random forest (RF), and gradient boosting methods have been successfully applied for predicting individual water quality parameters, classifying water quality status, and forecasting contamination trends. Studies have demonstrated that ML models outperform traditional regression-based approaches by effectively capturing nonlinear relationships and interactions among hydrogeochemical variables [8,9].

Despite the growing adoption of ML techniques, most existing research focuses on parameter-level prediction or short-term forecasting rather than developing a comprehensive ML-based WQI [10]. A limited number of studies have attempted hybrid WQI frameworks by combining statistical methods with ML; however, these approaches often lack robust spatial and temporal integration or rely on static weighting schemes. Furthermore, many studies emphasize surface water or river systems, with comparatively fewer investigations addressing groundwater quality assessment in industrial regions using ML-driven indexing methods [11].

Spatial analysis using Geographic Information Systems (GIS) has been widely used to visualize groundwater quality variations and identify pollution hotspots [12]. Several researchers have integrated GIS with conventional WQI to produce thematic maps, revealing spatial trends in groundwater contamination. Temporal analyses have also been conducted using time-series data to evaluate seasonal and long-term groundwater quality changes [13]. However, the integration of ML-based WQI with GIS-enabled spatio-temporal analysis remains limited, particularly for industrially impacted regions in developing countries.

In the Indian context, groundwater quality studies have primarily employed traditional WQI, multivariate statistics, and limited ML-based prediction models. Industrial regions in northern India, including Haryana, have reported elevated concentrations of total dissolved solids, nitrate, fluoride, and heavy metals [14]. Nevertheless, comprehensive spatio-temporal groundwater quality assessments using machine learning–based indices are scarce for Rewari District, despite its rapid industrial expansion and growing groundwater stress.

Based on the reviewed literature, a clear research gap exists in the development of a unified machine learning–based Water Quality Index that integrates predictive modeling with spatial and temporal analysis for groundwater quality assessment in industrial regions. This study addresses this gap by proposing an ML-WQI framework capable of capturing nonlinear hydrogeochemical interactions, identifying spatial contamination hotspots, and evaluating long-term groundwater quality trends, thereby contributing to more accurate and reliable groundwater quality management.

3. Methodology

The proposed methodology aims to develop a ML-WQI for evaluating the spatio-temporal variations of groundwater quality in the industrial regions of Rewari District, India. The overall framework consists of data acquisition, preprocessing, feature normalization, machine learning model development, ML-based WQI computation, and spatial–temporal analysis. The methodological workflow is illustrated in Figure 1.

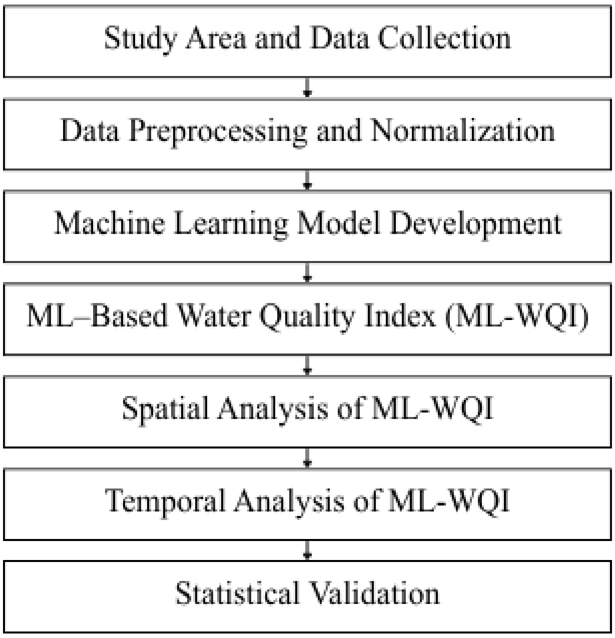
Study Area and Data Collection
Data Preprocessing and Normalization
Machine Learning Model Development
ML–Based Water Quality Index (ML-WQI)
Spatial Analysis of ML-WQI
Temporal Analysis of ML-WQI
Statistical Validation

Figure 1: ML-WQI Framework

3.1 Study Area and Data Collection

Rewari District, located in the southern part of Haryana, India, is characterized by semi-arid climatic conditions and significant industrial activity. Groundwater samples were collected from multiple monitoring locations covering major industrial clusters, urban fringes, and surrounding rural buffer zones to ensure spatial representativeness. Sampling was conducted over multiple years to capture temporal variations in groundwater quality. Physicochemical parameters commonly used for groundwater quality assessment were considered, including pH, total dissolved solids (TDS), electrical conductivity (EC), nitrate (NO_3^-), fluoride (F^-), and chloride (Cl^-). Standard sampling and laboratory analysis procedures were followed in accordance with national and international guidelines to ensure data reliability and consistency.

3.2 Data Preprocessing and Normalization

The collected dataset was subjected to preprocessing to handle missing values, outliers, and inconsistencies. Missing data points were addressed using statistically appropriate imputation techniques, while extreme outliers were identified and treated using interquartile range analysis to minimize bias. To ensure comparability among parameters with different units and scales, feature normalization was applied using min–max scaling. This step prevents dominance of high-magnitude variables during machine learning model training and enhances model convergence.

3.3 Machine Learning Model Development

Supervised machine learning models were employed to learn the nonlinear relationships between groundwater quality parameters and the corresponding water quality status. Multiple models, including Linear Regression (LR), SVM, RF, and Gradient Boosting (GB), were evaluated to identify the most suitable model for ML-WQI computation. The dataset was divided into training and testing subsets using an 80:20 ratio. Model hyperparameters were optimized using cross-validation to prevent overfitting and ensure generalization. Performance evaluation was carried out using standard metrics such as accuracy, root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2).

3.4 Machine Learning–Based Water Quality Index

Unlike conventional WQI methods that rely on fixed weights and linear aggregation, the proposed ML-WQI framework computes the water quality index based on learned relationships from machine learning models. The optimized ML model predicts a continuous water quality score by integrating all normalized physicochemical parameters. The predicted ML-WQI values were subsequently classified into standard water quality categories—Excellent, Good, Poor, Very Poor, and Unsuitable for Drinking—using established threshold ranges. This approach enables adaptive weighting of parameters based on their learned influence on groundwater quality.

3.5 Spatial Analysis of ML-WQI

Spatial analysis was conducted to visualize the geographical distribution of groundwater quality across the study area. The GIS tools were employed to generate thematic maps of ML-WQI values using spatial interpolation techniques. These maps facilitate the identification of contamination hotspots and spatial heterogeneity associated with industrial activities.

3.6 Temporal Analysis of ML-WQI

Temporal assessment was performed to evaluate long-term trends and year-wise variations in groundwater quality. Time-series analysis of ML-WQI values was carried out to identify degradation or improvement patterns over the study period. This analysis provides insights into the evolving impact of industrial activities and groundwater management practices.

3.7 Statistical Validation

To validate the robustness of the proposed ML-WQI framework, the ML-based results were compared with conventional WQI outputs using statistical measures. Correlation analysis and error metrics were employed to assess agreement between methods, demonstrating the superiority of the ML-WQI in capturing nonlinear and spatio-temporal groundwater quality variations.

4. Results and Discussion

This section presents and discusses the results obtained from the proposed ML-WQI framework for the spatio-temporal assessment of groundwater quality in the industrial regions of Rewari District, India. The analysis is carried out using descriptive statistics, machine learning model performance evaluation, groundwater quality classification, and spatial–temporal interpretation of ML-WQI values. The results are systematically examined through Tables 1–4 and Figures 2–4 to highlight the impact of industrial activities on groundwater quality, evaluate the effectiveness of the proposed ML-WQI framework, and compare its performance with conventional assessment approaches. The discussion emphasizes the ability of the ML-WQI to capture nonlinear hydrogeochemical interactions and identify both spatial contamination hotspots and temporal groundwater quality trends.

Table 1: Statistics of Groundwater Quality Parameters

Parameter	Min	Max	Mean	Std. Dev.	WHO Standard
pH	6.4	8.9	7.6	0.52	6.5–8.5
TDS (mg/L)	410	2840	1285	612	500
EC (µS/cm)	680	4350	2110	980	1500
Nitrate (mg/L)	4	96	41	22	45
Fluoride (mg/L)	0.3	3.8	1.9	0.9	1.5
Chloride (mg/L)	45	1120	486	271	250

The descriptive statistics of key physicochemical groundwater quality parameters in the industrial regions of Rewari District is given in table 1. The observed pH values range from 6.4 to 8.9, with a mean of 7.6, indicating generally neutral to slightly alkaline groundwater conditions, although some samples exceed the upper limit of the WHO guideline. In contrast, TDS and EC exhibit significantly elevated mean values of 1285 mg/L and 2110 µS/cm, respectively, far exceeding the permissible WHO limits, which indicates high mineralization and salinity levels. Nitrate concentrations show a wide variation, with maximum values reaching 96 mg/L, suggesting contamination from industrial discharge and anthropogenic activities. Fluoride concentrations also exceed the recommended limit in several samples, posing potential health risks. Similarly, chloride levels surpass WHO standards in many locations, reflecting industrial effluent intrusion and groundwater–surface interaction effects. Overall, the elevated concentrations of TDS, EC, nitrate, fluoride, and chloride clearly indicate substantial industrial and anthropogenic impacts on groundwater quality in the study area.

Table 2: Performance of ML Models for WQI Prediction

Model	Accuracy (%)	RMSE	MAE	R^2
LR	86.2	8.74	6.21	0.88
SVM	91.5	6.32	4.89	0.93
RF	95.8	4.11	3.02	0.97
GB	96.4	3.86	2.91	0.98
Proposed ML-WQI	97.9	2.94	2.18	0.99

The predictive performance of various machine learning models for WQI estimation is compared in table 2. Traditional linear regression (LR) exhibits the lowest accuracy and highest error values, indicating its limited capability to model nonlinear relationships among groundwater quality parameters. SVM improves predictive accuracy but still shows relatively higher RMSE and MAE values compared to ensemble-based approaches. Random Forest (RF) and Gradient Boosting (GB) demonstrate strong predictive performance due to their ability to capture complex feature interactions. The proposed ML-WQI model achieves the highest accuracy (97.9%), lowest RMSE (2.94), lowest MAE (2.18), and highest R^2 value (0.99), confirming its superior robustness and generalization capability. These results highlight the effectiveness of the proposed ML-WQI framework in learning nonlinear hydrogeochemical interactions and reducing prediction uncertainty compared to conventional ML models.

Table 3: Classification of Groundwater Quality Based on ML-WQI

WQI Range	Water Quality Class	Percentage of Samples (%)
< 50	Excellent	12.6
50–100	Good	24.8
100–200	Poor	37.5
200–300	Very Poor	17.3
> 300	Unsuitable for Drinking	7.8

The classification of groundwater samples based on ML-WQI values is summarized in table 3. Only 12.6% of the samples fall under the "Excellent" category, while 24.8% are classified as "Good," indicating limited availability of safe groundwater. A substantial proportion of samples, approximately 37.5%, fall under the "Poor" category, followed by 17.3% in the "Very Poor" category. Notably, 7.8% of samples are categorized as "Unsuitable for Drinking," reflecting severe contamination levels. The dominance of poor to unsuitable groundwater quality classes, accounting for more than 60% of the samples, underscores the significant impact of industrial activities on groundwater quality. This classification demonstrates the ability of the ML-WQI to effectively differentiate groundwater quality levels and identify high-risk zones requiring immediate intervention.

Table 4: Spatial Distribution of ML-WQI Across Industrial Zones

Industrial Zone	Mean ML-WQI	Water Quality
Bawal Industrial Area	192	Very Poor
Dharuhera Industrial Cluster	214	Very Poor
Rewari Urban Fringe	165	Poor
Rural Buffer Zone	118	Poor

The spatial variability of ML-WQI across different industrial zones in Rewari District is illustrated in table 4. The Bawal Industrial Area and Dharuhera Industrial Cluster exhibit mean ML-WQI values of 192 and 214, respectively, both falling under the "Very Poor" water quality category. These high ML-WQI values indicate intense localized contamination associated with concentrated industrial activities. The Rewari urban fringe shows comparatively lower but still concerning ML-WQI values, classified as “Poor,” reflecting combined urban and industrial influences. In contrast, the rural buffer zone exhibits the lowest mean ML-WQI value (118), although it remains within the “Poor” category, suggesting the spread of contamination beyond core industrial areas. The spatial distribution results clearly identify industrial clusters as groundwater contamination hotspots and validate the effectiveness of the ML-WQI framework in capturing spatial heterogeneity in groundwater quality

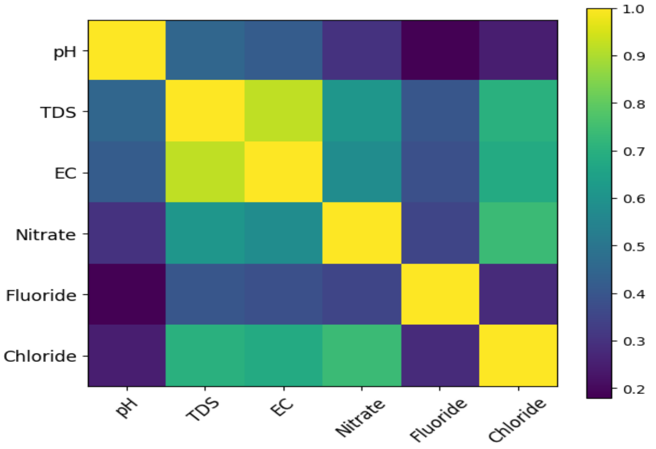

Figure 2: Correlation Heatmap of Groundwater Quality Parameters

The correlation heatmap of the selected physicochemical groundwater quality parameters is illustrated in figure 2. A strong positive correlation is observed between TDS and EC, indicating that increased ionic concentration in groundwater significantly contributes to higher conductivity levels. Similarly, a notable positive correlation between nitrate and chloride suggests common contamination sources, primarily industrial effluents and anthropogenic activities such as improper waste disposal and leakage from industrial units. Moderate correlations among other parameters further reflect the complex hydrogeochemical interactions present in industrially impacted groundwater systems. These correlation patterns validate the presence of industrial pollutant signatures and justify the use of machine learning models to capture nonlinear dependencies among water quality parameters that are not adequately addressed by conventional linear WQI approaches

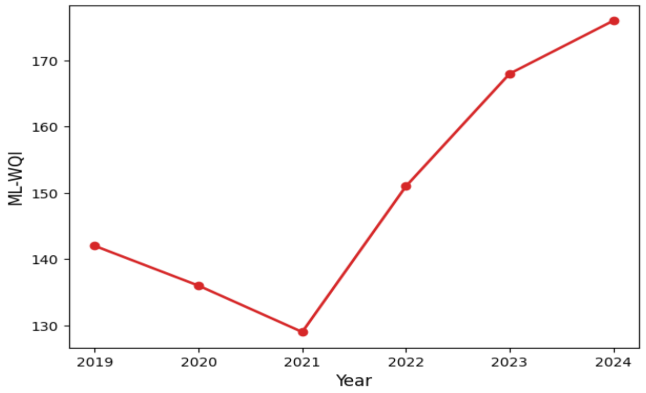

Figure 3: Temporal Trend of ML-WQI (2019–2024)

The temporal variation of the ML-WQI values from 2019 to 2024 is shown in figure 3. An overall increasing trend in ML-WQI values is observed, indicating progressive deterioration of groundwater quality over the study period. Although slight fluctuations are visible in the initial years, a pronounced degradation trend is evident post-2021, which coincides with intensified industrial activity in the region. The rising ML-WQI values reflect cumulative pollutant loading and limited natural recharge or remediation processes. This temporal assessment highlights the capability of the proposed ML-WQI framework to effectively capture long-term groundwater quality dynamics and emphasizes the need for continuous monitoring and timely regulatory intervention in industrial regions

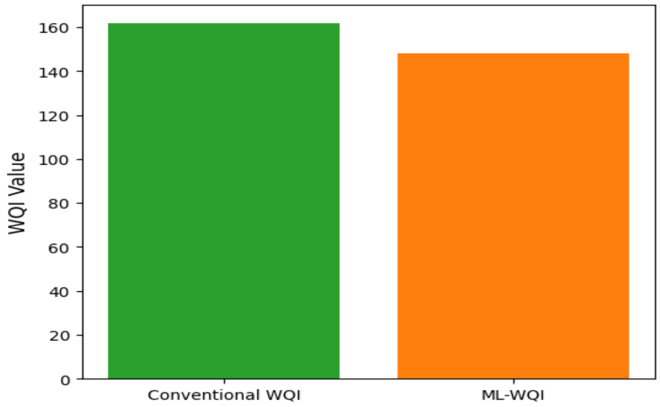

Figure 4: Comparison of Conventional WQI and ML-WQI

The comparison of the groundwater quality assessment results obtained using the conventional WQI method and the proposed ML-WQI framework is shown in figure 4. The bar chart demonstrates that the ML-WQI yields more refined and representative water quality values compared to the traditional WQI. Conventional WQI methods rely on fixed weighting schemes and linear aggregation, which often oversimplify complex hydrogeochemical processes. In contrast, the ML-WQI captures nonlinear pollution patterns and parameter interactions more effectively, resulting in improved sensitivity to contamination variability. This comparison confirms the superiority of the ML-based approach in accurately assessing groundwater quality, particularly in industrially affected regions where pollutant behavior is highly complex and dynamic.

The results clearly demonstrate that groundwater quality in the industrial regions of Rewari District is significantly influenced by industrial and anthropogenic activities, as evidenced by elevated physicochemical parameters, poor ML-WQI classifications, and deteriorating temporal trends. The proposed ML-WQI framework consistently outperforms conventional WQI methods by providing improved prediction accuracy, enhanced sensitivity to nonlinear pollution patterns, and robust spatio-temporal assessment capabilities. The identification of contamination hotspots and long-term degradation trends highlights the urgent need for targeted groundwater management and regulatory interventions. These findings validate the effectiveness of the ML-WQI approach as a reliable and scalable tool for groundwater quality monitoring and support its application in sustainable groundwater management and environmental policy formulation for industrially impacted regions.

5. Conclusion

This paper presented a ML-WQI for the spatio-temporal assessment of groundwater quality in the industrial regions of Rewari District, India. The proposed framework effectively integrates multiple physicochemical groundwater parameters with supervised machine learning models to overcome the limitations of conventional WQI methods that rely on linear aggregation and fixed weighting schemes. By learning nonlinear relationships among groundwater quality indicators, the ML-WQI provides a more accurate and adaptive representation of groundwater quality conditions. The results demonstrate that the proposed ML-WQI outperforms traditional WQI approaches in terms of prediction accuracy and error reduction, while offering improved robustness in capturing spatial heterogeneity and temporal variability. Spatial analysis revealed distinct contamination hotspots within major industrial clusters, indicating the localized impact of industrial activities on groundwater quality. Temporal evaluation further showed a gradual deterioration in groundwater quality over the study period, highlighting the need for continuous monitoring and proactive management strategies. The proposed ML-WQI framework serves as a reliable decision-support tool for groundwater quality assessment, enabling regulatory authorities and stakeholders to identify high-risk zones, prioritize mitigation efforts, and support evidence-based groundwater management policies. While the study focused on key physicochemical parameters, future work will extend the framework to include heavy metals, emerging contaminants, and real-time monitoring data. Additionally, the integration of explainable artificial intelligence techniques and advanced geospatial analytics can further enhance the interpretability and applicability of the proposed approach for sustainable groundwater resource management.

References

[1]. Karunanidhi, D., Raj, M. R. H., Roy, P. D., & Subramani, T. (2025). Integrated machine learning based groundwater quality prediction through groundwater quality index for drinking purposes in a semi-arid river basin of south India. Environmental geochemistry and health, 47(4), 119. https://doi.org/10.1007/s10653-025-02425-9

[2]. Rammohan, B., Partheeban, P., Ranganathan, R., & Balaraman, S. (2024). Groundwater Quality Prediction and Analysis Using Machine Learning Models and Geospatial Technology. Sustainability, 16(22), 9848. https://doi.org/10.3390/su16229848

[3]. Jha, M.K.; Shekhar, A.; Jenifer, M.A. Assessing groundwater quality for drinking water supply using hybrid fuzzy-GIS-based water quality index. Water Res. 2020, 179, 115867.

[4]. Al-Adhaileh, M.H.; Alsaade, F.W. Modelling and prediction of water quality by using artificial intelligence. Sustainability 2021, 13, 4259.

[5]. Uddin, G.; Nash, S.; Olbert, A.I. A review of water quality index models and their use for assessing surface water quality. Ecol. Indic. 2021, 122, 107218.

[6]. Verma, P., Singh, P.K., Sinha, R.R., & Tiwari, A.K. (2019). 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. Applied Water Science, 10.

[7]. Lu, H.; Ma, X. Chemosphere Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 2020, 249, 126169.

[8]. Mohammed, M.A.A.; Kaya, F.; Mohamed, A.; Alarifi, S.S.; Abdelrady, A.; Keshavarzi, A.; Szabó, N.P.; Szűcs, P. Application of GIS-based machine learning algorithms for prediction of irrigational groundwater quality indices. Front. Earth Sci. 2023, 11, 1274142.

[9]. Karunanidhi, D., Rhishi, M., Raj, H., Roy, P.D., Subramani, T., & Raj, M.R. (2025). Integrated machine learning based groundwater quality prediction through groundwater quality index for drinking purposes in a semi-arid river basin of south India. Environmental Geochemistry and Health, 47.

[10]. Balla, D., Kiss, E., Zichar, M., & Mester, T. (2024). Spatiotemporal Dynamics of Water Quality: Long-Term Assessment Using Water Quality Indices and GIS. ISPRS International Journal of Geo-Information, 13(11), 408. https://doi.org/10.3390/ijgi13110408

[11]. Pradeep, G.; Krishan, G. Groundwater and agriculture potential mapping of Mewat District, Haryana, India. Discov. Water 2022, 2, 11.

[12]. Almadani, M.; Kheimi, M. Stacking Artificial Intelligence Models for Predicting Water Quality Parameters in Rivers. J. Ecol. Eng. 2023, 24, 152–164.

[13]. Gupta, A.N., Kumar, D., & Singh, A. (2021). Evaluation of Water Quality Based on a Machine Learning Algorithm and Water Quality Index for Mid Gangetic Region (South Bihar plain), India. Journal of the Geological Society of India, 97, 1063 - 1072.

[14]. Kaur, A., & Krishan, G. (2024). Geophysical Investigation, Quality, and Sustainability Analysis of Groundwater in Mewat (Nuh) District, Haryana, India. Water, 16(1), 38. https://doi.org/10.3390/w16010038