Indian Journal of Engineering & Materials Sciences
Vol. 32, October 2025, pp. 536-551
DOI: 10.56042/ijems.v32i05.20867

Innovations in water quality management using machine learning approaches

Shivangi Bharadwaj*, Ashok Kumar Gupta, & Anil Kumar Sahu
Department of Civil Engineering, Delhi Technological University, New Delhi 110 042, India

Received: 21 June 2025; accepted: 29 August 2025

Groundwater contamination has been posing a significant threat to sustainable water resource management, particularly in industrialized and urbanized regions. This research has introduced a novel, data-driven framework that integrates machine learning, statistical data analysis, and feature optimization to evaluate and forecast groundwater quality. Analytical results of 488 groundwater samples had been tested, and four feature reduction scenarios had been implemented using Pearson correlation to evaluate predictive performance with minimal input variables. Statistical analysis has highlighted elevated levels of parameters such as Electrical Conductivity, Chloride, Magnesium, and Total Hardness, exceeding permissible limits, and have been causing most samples to be unsuitable for consumption without treatment. To enhance groundwater monitoring and reduce laboratory testing costs, six machine learning algorithms, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, XGBoost, and Artificial Neural Network, have been used to predict the Weighted Arithmetic Water Quality Index. Model accuracy had been tested using statistical metrics such as R^2, RMSE, MAE, MAPE, and CRMSE, with effectiveness assessed using Taylor diagrams. ANN exhibited the highest accuracy even when using a single input (K), while SVM maintained consistent reliability with only two inputs (Mg and K), providing a cost-effective monitoring solution. Validation with 70 independent datasets has confirmed the robustness and applicability of the suggested methodology. The study has presented an innovative modeling strategy that has substantially decreased laboratory testing needs while preserving predictive reliability. Additionally, it has offered practical implications for scalable, cost-effective deployment in areas with water scarcity or insufficient dataset.

Keywords: Groundwater contamination, K-fold cross validation, Machine learning models, Regional hydrology, Water quality index

1 Introduction

Groundwater constitutes approximately 98% of the global freshwater reserves, which are extensively used for household, agricultural, and industrial purposes. The chemical composition of groundwater is crucial for evaluating its suitability for human consumption[1-2]. Geological characteristics and environmental factors affect the quality of groundwater. Globally, groundwater contamination is a significant issue that presents grave risks to public health and environmental quality[3]. A Water quality index (WQI) is a quantitative measure used to assess water quality for various purposes. It can be used to evaluate the appropriateness of water for drinking, industrial uses, aquatic life, and other applications. Greater WQI values indicate superior water quality[4]. Water Quality Classification (WQC) is a system that classifies water as either mildly contaminated or pure based on the WQI value range[5]. Agriculture serves as the primary consumer of water resources[6-7]. The over use of groundwater in areas with significant climate variability has led to the deterioration of this resource in terms of both physical and chemical properties.

Purified water is an essential resource upon which living creatures depend. Consequently, the development of a water quality forecasting technique to predict future water quality conditions has great social and economic importance. Predictive models allow for real-time or near-real-time estimation of the WQI and WQC, making them more efficient and cost-effective than traditional laboratory analyses. This capability enables continuous monitoring of water quality, early identification of declines, and prompt responses to potential hazards or pollution events.

A number of researchers have assessed the potability of groundwater using the following major parameters, pH, Total Hardness (TH), Electrical Conductivity (EC), Magnesium (Mg^{2+}), calcium (Ca^{2+}), sodium (Na^{+}), potassium (K^{+}), bicarbonate (HCO_3^-), nitrate (NO_3^{2-}), fluoride (F), and sulphate (SO_4^{2-})[8-11]. Because measuring any of these factors in groundwater is a time-consuming and costly process,

*Corresponding author
(Email: shivangibharadwaj_2k21phdce03@dtu.ac.in)

it can be challenging.Therefore, it is very difficult to evaluate water quality while lowering the effective cost and subjectivity. Numerous water quality indices, which are available in the literature, have been developed in the past based on different water quality criteria in consideration of the common problems associated with the straight forward determination of water quality.

As machine learning techniques have performed better than statistical methods over the past three decades, they have been widely utilized to estimate groundwater quality for drinking [12-14]. As a result, the researchers have discussed index-based techniques for assessing the quality of water suitable for human consumption, such as the US National Sanitation Foundation WQI[15], the Canadian WQI[16]. WQI and Equivalent WQI, which is thought to be a useful technique for giving precise and thorough information about the general quality of water for drinking purposes.

Computational intelligence approaches, including artificial neural networks and genetic algorithms, are increasingly focused on predicting environmental changes over time through time-series analysis for water quality monitoring and forecasting. These methods are particularly valuable because they enable the modelling of nonlinear systems and are resistant to noise data, resulting in more accurate outcomes[17-19]. Thus, machine learning reduces the time spent on computation needed to determine the water quality parameter for every specimen. Calculating the Water quality index for a hundred tests using conventional formulas is time-consuming, whereas leveraging machine learning techniques is efficient[17-20].

This study examines 488 groundwater samples, using seven key hydro-chemical parameters,pH, Electrical Conductivity (EC), Chloride (Cl), Magnesium (Mg), Calcium (Ca), Potassium (K), and Total Hardness (TH) to assess water quality through the Weighted Arithmetic Water Quality Index (WAWQI) method. The main objective is to minimize the amount of input parameters necessary for effective groundwater quality evaluation.To improve efficiency, various models were evaluated to determine their ability to minimize the number of parameters required for WQI calculation, ultimately aiming to reduce the time and cost associated with traditional laboratory testing. Some studies also highlighted the optimization of input features to improve model interpretability and efficiency in practical applications[21].The performance of the leading models will be validated using an independent dataset of 70 samples, with predictions analyzed using Root Mean Square Error (RMSE) to guarantee robustness and reliability.

Previous studies have predominantly concentrated on forecasting the Water Quality Index (WQI) without assessing the significance of individual parameters, relied on single train-test data splits, and exhibited insufficient model validation. This research introduces a novel framework that systematically identifies and minimizes essential input parameters for accurate WQI estimation. The impact of diverse water quality indicators on the WQI will be examined utilizing Pearson correlation coefficients and pair plots, while geospatial maps generated through Inverse Distance Weighting (IDW) will assist in identifying areas with varying water quality conditions. To improve predictive accuracy, six machine learning models, including KNN, SVM, DT, RF, XGBoost, and ANN, will be optimized by nested 5-fold cross validation and grid search. Unlike earlier models depending on complete parameter sets and restricted validation, this study validates performance using an independent dataset and shows that parameter-efficient models can outperform WQI prediction and spatial risk mapping. This research contributes a strategically optimized framework for intelligent water resource management, aligning with both national and global sustainability goals.

Future study can leverage these created AI models, which can be expanded to include other characteristics such as biological and heavy metal pollution in groundwater for calculating WQI. Furthermore, the proposed model can be tailored to other geographical regions. These ML models can be combined with real-time sensor data to provide continuous monitoring of GW.

2 Materials and Methods

2.1 Studyarea

The study was performed in Delhi and the National Capital Region (NCR) of India, located between 26.8042°N to 30.1380°N latitude and 75.5158°E to 78.2567°E longitude as shown in Fig. 1. Accelerated urbanization and unsustainable groundwater withdrawal have resulted in significant water scarcity, rendering the region vulnerable to land subsidence. The Ganga and Yamuna rivers, two of India's most important waterways, traverse this region,

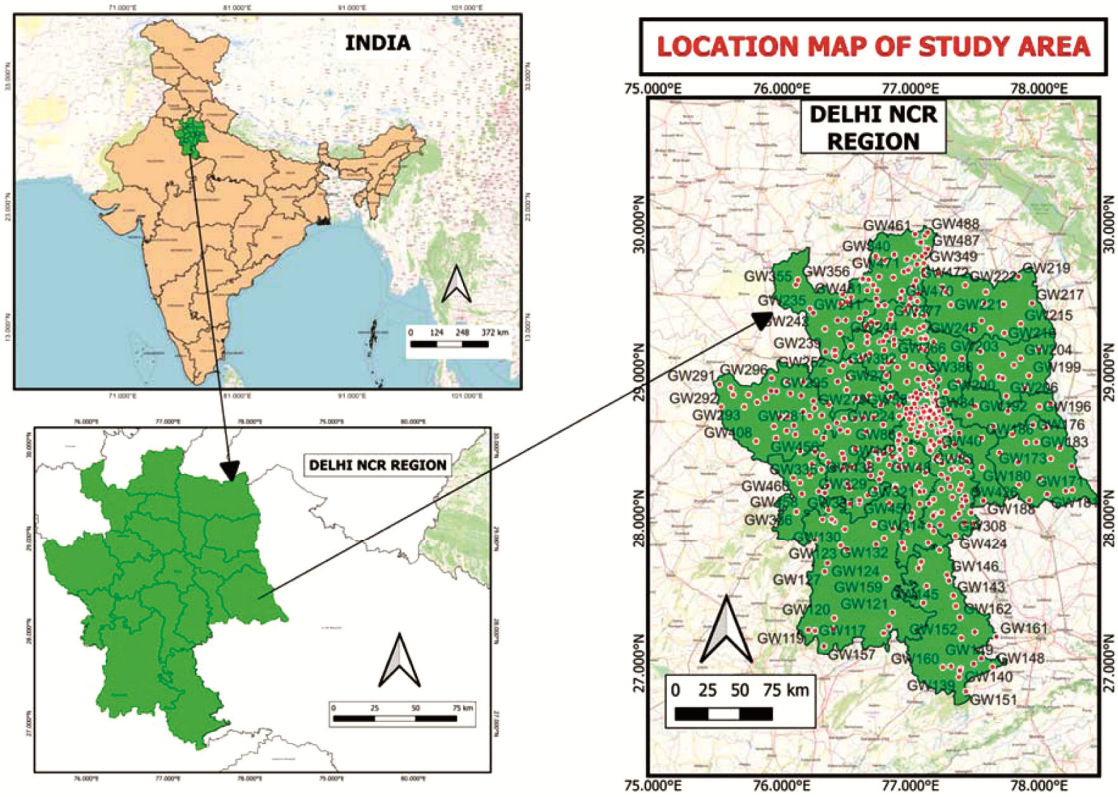

Fig. 1 — GW sample location from Delhi NCR, India.

significantly contributing to its hydrological and biological equilibrium. The geology of thisarea is located at the confluence of the Indo-Gangetic Alluvial Plains to the north and east and the Aravalli Range in the central, southern, and western areas. This region consists of hard rock formations (ridges) and the Quaternary Alluvium Plain.

2.2 Groundwater samples

This analysis focuses on seven key parameters: pH,EC, Cl,Mg,Ca,K, and TH. The research employsa dataset of 36 groundwater samples collected from various locations during November and January 2023-2024near the Bhalswa dump in Delhi, India, extending to Delhi Technological University in Rohini, Delhi. The samples were collected by boreholes, hand pumps, and submersible pumps. The laboratory analysis and collection of GW samples are illustrated in Fig. 2(a, b, c). Furthermore, an augmented dataset of 452 groundwater records was acquired from the Central Ground Water Board (CGWB) in 2021[22].

2.3 Weighted arithmetic water quality index (WAWQI)

WQI is a crucial instrument that consolidates extensive data on multiple parameters, facilitating the

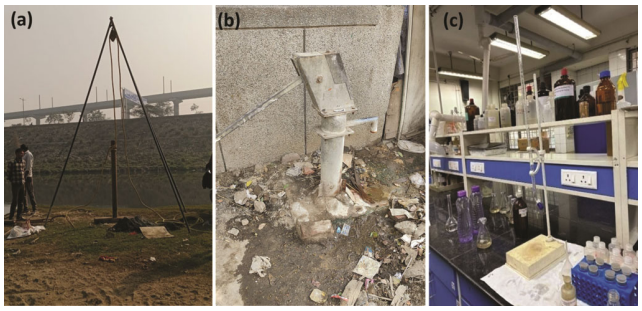

Fig. 2 — (a) Collection of groundwater samples using Borehole, (b) GW Sample collection using Handpump, and (c) Analysis of GW samples in laboratory.

reduction of water quality information into a singular value. The values that are acquired can be classified into various categories that signify the quality of water. The utilization of these indices provides a comprehensive assessment of water quality conditions and determines the suitability of water for specific applications such as irrigation and drinking water supply[23]. Numerous water quality indices exist, such as National Sanitation Foundation Water Quality Index (NSFWQI), Canadian Council of Ministers of Environment Water Quality Index (CCMEWQI), Oregon Water Quality Index (OWQI), and Weight Arithmetic Water Quality Index (WAWQI)[24].

No singular Water Quality Index exists for evaluating surface water quality, but numerous modifications have been proposed to create different WQIs suited to specific regional conditions[25]. The flowchart in Fig. 3 outlines stepwise process for calculating an overall index, starting with the selection of key parameters, followed by the creation of sub-indices, assigning weights to parameters, and aggregating sub-indices to obtain the final water quality index. Table 1 illustrates sample dataset parameters and WQI.

The Weighted Arithmetic Water Quality Index (WAWQI) is a method used to assess overall water quality by considering multiple physicochemical parameters. Each parameter is assigned a weight based on its relative significance to water quality, and the index is calculated using the following formula from equations(1-4)[26-27].

$$Q_i = [\frac{V_a - V_i}{V_s - V_i}] \times 100 \quad \ldots(1)$$

$$K = \frac{1}{\sum(\frac{1}{V_S})} \quad \ldots(2)$$

$$W_i = \frac{K}{V_S} \quad \ldots(3)$$

$$WQI = \frac{\sum Q_i W_i}{\sum W_i} \quad \ldots(4)$$

Where, Q = Quality rating scalefor each parameter
V_a = Actual value of water quality parameter obtained from the data
V_i = Ideal value of water quality parameter (pH=7, all other parameters, it is zero
V_s = Recommended standard value of the parameter
K = Proportionality constant
Wi= Unit weight for each parameter
WQI = Water Quality Index

2.4 Machine learning models

Machine learning is a branch of artificial intelligence which teaches machines to forecast results using previous data and experiences; while artificial intelligence is the overall field that seeks to simulate human capabilities[28]. Machine learning allows systems to improve and develop from experience on their own without requiring explicit programming[29]. Machine learning has been extensively employed as an effective instrument for addressing groundwater environmental challenges, as it can forecast water quality, optimize resource allocation, and mitigate water scarcity difficulties[30]. This study established six machine learning models, including Decision Tree, Artificial Neural Network, Random Forest Regressor, Support Vector Machine, XGBoost, and K-Nearest Neighbors, to predict WA-WQI using a supervised regression methodology for estimating the Water Quality Index. Models were chosen based on their efficacy, ease of use, and demonstrated application in predicting groundwater quality.

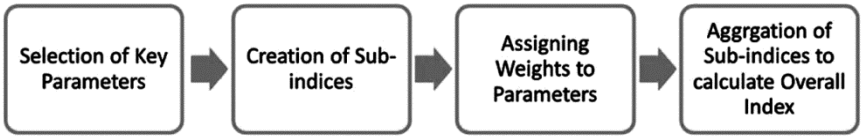

Fig. 3 — Steps for developing WA-WQI.

Table 1 — Sample dataset illustrating essential parameters and WQI

Sample No.	pH	EC	Cl	Mg	Ca	K	TH	WQI
1	8.1	2165	389	83	112	178	620	566.663
2	8.41	3019	708	58	44	5.7	350	94.5349
3	7.99	4808	1098	202	128	14	1151	182.527
4	8.39	623	56	22	52	4.1	220	69.1221
5	7.64	13060	3871	618	400	32	3542	460.138
6	7.9	4823	1273	229	216	9.2	1481	188.871
7	8.43	707	118	39	16	8.8	200	88.0955
8	8.14	2309	298	48	40	8.5	300	86.1793
9	8.82	890	90	36	20	1.6	200	79.3606
10	8.13	5480	637	129	68	14	701	143.89
11	7.5	8243	2779	80	801	9	2332	170.042
12	8.22	802	106	41	12	1.8	200	63.2956
13	7.51	21200	4697	846	505	23	2371	531.933
14	8.94	1196	55	19	12	3.8	110	79.9898
15	8.85	1190	111	34	12	98	170	339.152

2.4.1 Decision Tree (DT)

A Decision Tree (DT) is a non-parametric model, indicating that it does not presume any specific distribution of the data. The process involves recursively splitting the dataset at each node into two child nodes according to feature values, with the objective of minimizing the Mean Squared Error (MSE) of the target variable[31-32]. The procedure persists until a stopping requirement, such as maximum tree depth or minimum sample size per node, is satisfied. For novel predictions, the model allocates the mean target value of the leaf node corresponding to the input data, which constitutes the ultimate predicted result[33].

2.4.2 Artificial neural network (ANN)

ANN is a machine learning model inspired by the information processing mechanisms of biological neural networks in the human brain[34]. The Multilayer Perceptron is a commonly utilized artificial neural network employed for classification and regression tasks[35]. It is made up of multiple layers, including an input layer, one or more hidden layers, and an output layer, with each layer containing perceptive units known as neurons[36].

2.4.3 Random forest (RF)

Ensemble methods enhance the predictive accuracy of an individual model by training several models and merging their predictions[37]. These methods are regarded as leading solutions for numerous problems in machine learning[38]. RF computes the average of the predictions from all trees to produce the final output in case of regression[39]. While individual decision trees are straightforward to interpret, this clarity is lost in random forests due to the aggregation of multiple trees in exchange for better performance[40]. Random Forest combines multiple decision trees as base models by integrating bagging technique and random feature selection. Bagging (bootstrap aggregating) randomly resamples the original data to create a training dataset with replacements, known as bootstrap samples[41]. Each tree in RF is trained on a distinct bootstrap sample from the training set by using a randomly selected subset of the predictor (feature) variables[42-43].

2.4.4 Support vector machine (SVM)

SVM was introduced to the computer science field by Vladimir Vapnik in 1995. It is a widely used and highly effective supervised machine learning algorithm[44].

2.4.5 XGBoost (XGB)

XGBoost is Extreme Gradient Boosting became popular in ML competitions after being developed in multiple languages like Python, R, and Java[45]. It's essentially an advanced ensemble of decision trees with a modified gradient boosting framework[46].

2.4.6 K- Nearest neighbors (KNN)

It was first developed by Evelyn Fix and Joseph Hodges in 1951 as part of military research[47]. KNN regression is a lazy and non-parametric algorithm, as it does not need a pre-built model for new data points[48-49]. It is one of the simplest AI calculations to grasp since it just stores all the data at hand and organizes additional information based on similarity[50].

2.4.7 Resampling techniques

This research uses 5-fold cross-validation to evaluate machine learning models. The dataset is divided into five subsets, with the model trained on four and evaluated on one in each iteration. The process repeats for all folds, and average performance metrics provide a reliable evaluation. This technique optimizes data usage by ensuring each instance serves for both training and testing.

2.5 Model evaluation metrics

Performance evaluation metrics are essential for assessing machine learning prediction models, providing a mathematical basis to compare predicted and actual values. In this study, coefficient of determination (R^2) and root mean square error (RMSE) were utilized to analyze model accuracy. These metrics were further employed to construct a Taylor diagram, enabling the identification of the best-performing models.

2.5.1 Coefficient of determination (R^2 or R-squared)

R2 is the quotient or ratio of the explained variation compared to the total variation[51]. It is a dimensionless quantity.

2.5.2 Mean absolute error (MAE)

MAE is a calculation of the precise discrepancies between predicted and actual values across the test dataset in order to quantify the average error, but measures like the mean absolute error are easier to understand than MSE or RMSE[52]. It is expressed in the same unit as that of target variable.

2.5.3 Mean absolute percentage error (MAPE)

MAPE is a popular measure of forecast accuracy due to its scale independence and clarity[53].

2.6 Model's accuracy assessment using Taylor diagram technique

This study utilized Taylor diagrams to assess the precision of machine learning models in forecasting the Water Quality Index (WQI). Taylor Diagram is a polar graph that effectively showcases the model's performance relative to observed data using three statistical metrics: standard deviation, Pearson's correlation coefficient, and centered root mean squared (CRMSE) difference or error[54]. Evaluating the distance between each model and the reference point assesses the performance of the models. When the model point closely matches the reference point, it signifies similarity in standard deviation, elevated correlation, and a CRMSE near zero[55].

$$Centered\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(a_i - \bar{a}) - (b_i - \bar{b})]^2} \quad \ldots(5)$$

a_i is the i-th actual value
b_i is the i-th predicted value
$\bar{a}$ is the mean of the actual values
$\bar{b}$ is the mean of the predicted values
N is the number of observations

2.7 Data analysis

2.7.1 Pearson correlation coefficient

The Pearson correlation coefficient, denoted as 'r', is a widely used statistical metric that assesses the strength and direction of the linear relationship between two continuous variables[56.] The correlation strength increases with the absolute value of the correlation coefficient r. The values of r are confined to the interval [-1,1]. When r > 0, the variables exhibit a positive correlation. When r < 0, a negative correlation is observed. If r = 0, there exists no linear correlation among the variables[57].

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \quad \ldots(6)$$

X_i and Y_i are the values of X and Y for the i-th individual.
$\bar{X}$ is the mean of X.
$\bar{Y}$ is the mean of Y.
N is the number of data points.

Feature selection involves identifying a subset of the initial input features that show a strong correlation with the system output[58]. The literature has introduced numerous types of FSM, generally classified as filter, wrapping, and embedded methods[59]. Methods like the Pearson Correlation Coefficient are typically preferred for feature selection due to their speed, ease of use and automation, as well as their lower computing requirements. Moreover, their independence from the modeling framework ensures they remain impartial and may be utilized with datasets containing numerous predictor variables[60].

3 Results and Discussion

3.1 Water quality analysis

The Water Quality Index (WQI) is an essential instrument for evaluating water appropriateness for diverse purposes, such as consumption, irrigation, and industrial usage. Based on the analysis of 488 groundwater samples, the WQI classification[61] given by reveals significant water quality concerns. Only one sample falls under the "Excellent" category (WQI 0-25), making it safe for all uses. Thirteen samples are rated "Good" (WQI 25-50), suitable for drinking, irrigation, and industrial purposes. A significant number of samples (130) are classified as "Poor" (WQI 50-75), signifying that the water is unsuitable for consumption but may still be utilized for irrigation and industrial purposes. The highest number of samples, 224, belong to the "Very Poor" category (WQI 75-100), restricting their application solely to irrigation. Alarmingly, 120 samples exceed a WQI of 100, making them unsuitable for any use without proper treatment. These findings highlight the urgent need for groundwater monitoring and treatment strategies to improve water quality and ensure safe usage[62].

3.2 Statistical analysis

The statistical summary of the gathered water quality parameter data and the derived water quality index shown in Table 2. All the calculated values have been compared with different standards, i.e., IS-10500 (2012) code, European Union (Drinking Water) Regulations 2023, Water Quality Regulations 2021 (Fifth Edition), Abu Dhabi. The parameters Hardness (488.895 ±553.452 mg/L), Cl (407.261 ±686.812 mg/L), Mg (98.411 ±112.156 mg/L), and K (20.340 ±50.745 mg/L) exceed the BIS 10500:2012 values for drinking water. However, EC (2307.563 ±2716.073 µg/cm) and Ca (62.420 ±87.886 mg/L) are below the standard limit. The average pH value in the area is 8.277 ±0.467, suggesting that the majority of

Table 2 — Statistical summary of the dataset

Parameter	Median	SD	Min.	Mean	Max.	Permissible limit
TH	290	553.452	33	488.895	3814	200
K	5.95	50.745	0	20.34	520	12
Ca	32	87.886	4.4	62.42	801	75
Mg	49	122.156	2	98.411	846	30
EC	1397.5	2716.073	39.08	2307.56	24060	2500
pH	8.34	0.467	6.9	8.277	9.24	6.5-8.5
Cl	145.58	686.812	7.1	407.261	5388	250

- All values, with the exception of pH, EC, and WQI, are in mg/L
- pH, WQI are without dimension whereas EC (μS/cm)s
- All standards are from IS- 10500 (2012) code, except EC* and K**

* European Union (Drinking Water) Regulations 2023

** Water Quality Regulations 2021 (Fifth Edition), Abu Dhabi

the groundwater is alkaline. Total hardness refers to the concentration of calcium and magnesium ions, which can influence the taste of water and lead to scale buildup. High levels of chloride may indicate pollution or increased salinity. While magnesium is an essential mineral, excessive amounts can impact taste and health. Potassium is generally beneficial in small quantities, but high levels can be detrimental. The average and standard deviation of the WA-WQI for the basin are 148.241 ± 163.555.

3.3 Pair plot

A pair plot is a visualization technique that presents Kernel Density Estimation (KDE) plots for individual variable distributions on the diagonal and scatter plots for variable pairs off-diagonal, revealing both linear and non-linear relationships, as well as patterns and outliers in the dataset. Pair plots can help reveal clusters or groupings in the data, which is advantageous for unsupervised learning tasks. By visualizing these relationships, you can determine which features may have a strong impact on the target variable. Figure 4 shows the pair plot of all 7 parameters used for evaluating WQI. The study reveals a positive correlation between EC and Cl, Mg, and TH, suggesting that higher conductivity often correlates with higher levels of these elements. However, the data shows a skewed distribution of EC. The research indicates a robust positive link between chloride and EC and TH, a marginally positive linear association with Mg and Ca, and a skewed distribution of chloride data points. The scatter plots show a clear positive correlation between Mg and TH, a moderate positive correlation between Mg and WQI, and a skewed distribution of magnesium. The study demonstrates a robust positive link between TH and EC and Cl, as well as a distinct positive correlation between TH and Mg and Ca.

3.4 Geospatial distribution maps

The spatial distribution of the Water Quality Index (WQI) across Delhi NCR, India, is depicted in Fig. 5. The map illustrates varying water quality levels: green areas (WQI ~ 20) indicate relatively clean water with minimal contamination, while red regions (WQI ~ 1650) represent heavily polluted water, making it unfit for direct consumption. Yellow-orange zones signify moderate water quality, necessitating some degree of treatment. The northern and central parts of Delhi NCR exhibit severe pollution, as reflected by extensive red zones, whereas the southern and peripheral areas display better water quality with green patches. The gradual transition from green to red suggests increasing contamination, likely driven by urban expansion, industrial discharge, and agricultural runoff. Elevated WQI values in urban areas point to significant anthropogenic pollution sources such as sewage disposal, industrial effluents, and excessive groundwater extraction. Overall, the widespread poor water quality underscores the urgent need for groundwater conservation and effective remediation strategies.

3.5 Analysis of model performance utilizing all 7 features based on Pearson correlation coefficient to predict WQI

This study employs key water quality parameters 'pH,' 'EC,' 'Cl,' 'Mg,' 'K,' 'TH,' and 'Ca,' as features to predict the Water Quality Index (WQI) using the Nested K-Fold Cross Validation technique (k = 5) to enhance model reliability. Six machine learning models KNN, SVM, DT, RFR, XGBoost, and ANN as shown in Fig. 6. were evaluated using R^2, MAE, MAPE, RMSE, and Centered RMSE. ANN and SVM outperformed other models, achieving the highest R^2 (0.99), lowest MAE (4.518 for ANN), and minimal RMSE (8.294 for ANN, 9.689 for SVM), indicating superior predictive accuracy. XGBoost and Random

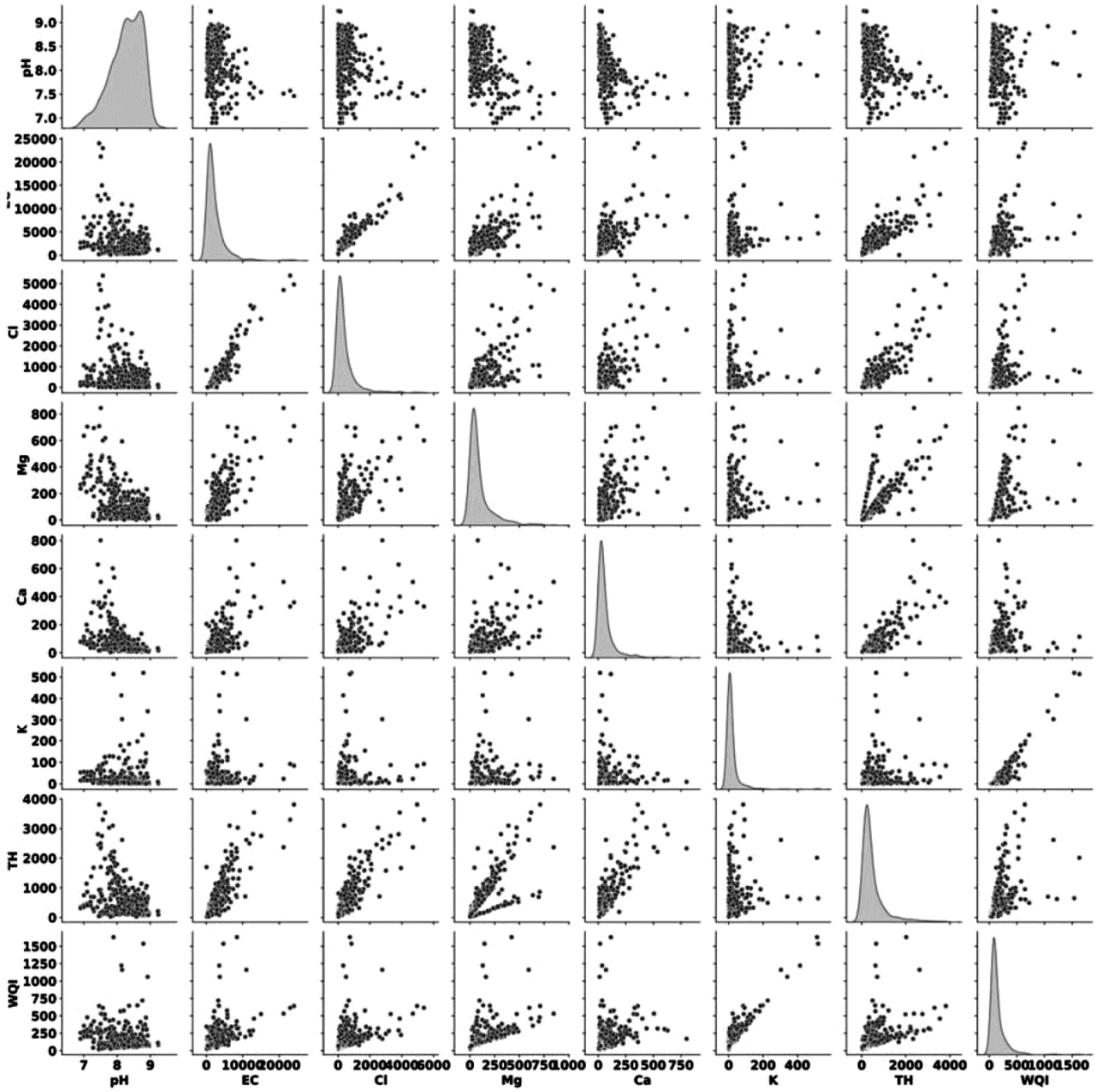

Fig. 4 — Pair plot representing all physicochemical characteristics of groundwater with water quality index.

Forest also demonstrated strong performance, with relatively low errors. Additionally, model ranking based on distance from the reference point in the Taylor diagram confirms ANN as the most accurate model (distance = 14.683), followed by SVM (15.146) and XGBoost (32.001). Random Forest, KNN, and Decision Tree showed comparatively lower accuracy, with DT ranking the lowest (57.001). The results highlight ANN and SVM as the most effective models for WQI prediction, offering high accuracy and minimal prediction errors. Based on this ANN has chosen for further validation.

3.6 Analysis of model performance using top 4 features based on Pearson correlation coefficient to predict WQI

The features used to predict WQI are EC, Mg, K, TH, and the target variable is WQI. The Nested K-Fold Cross Validation technique is used with k = 5. Table 3 shows the model architecture obtained via hyperparameter tuning for top 4 features.

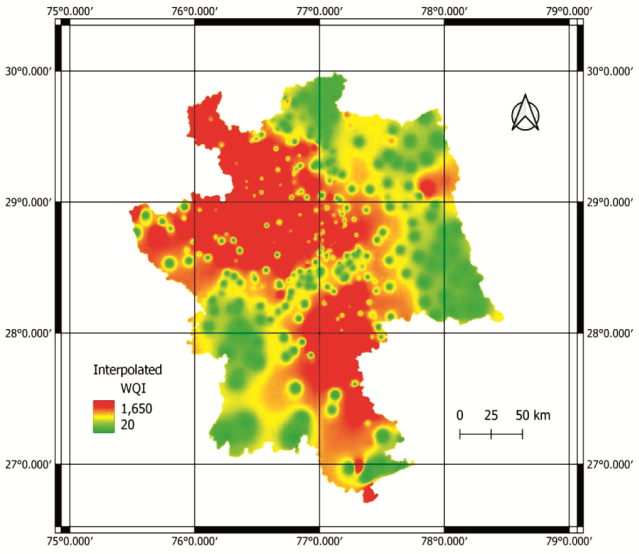

Fig. 5 — Spatial distribution of water quality index (WQI) in Delhi NCR, India.

Figure 7 (a-f) shows the line plot of actual vs. predicted WQI for KNN, SVM, DT, RF, XGBoost, ANN models for top 4 features.

3.6.1 Analysis of model accuracy for top 4 features utilizing a Taylor diagram

The Taylor diagram in Fig. 8 provides a visual representation of the performance of machine learning models using the top four features. Based on the rankings in Table 4, which evaluates models by their distances from the reference point in the Taylordiagram, the Support Vector Machine (SVM) model is identified as the most suitable. Consequently, this model is selected for further validation.

3.7 Analysis of model performance using top 2 features based on Pearson correlation coefficient to predict WQI

Mg and K are the selected features for predicting WQI as the target variable. The Nested K-Fold Cross-Validation technique is applied with k = 5

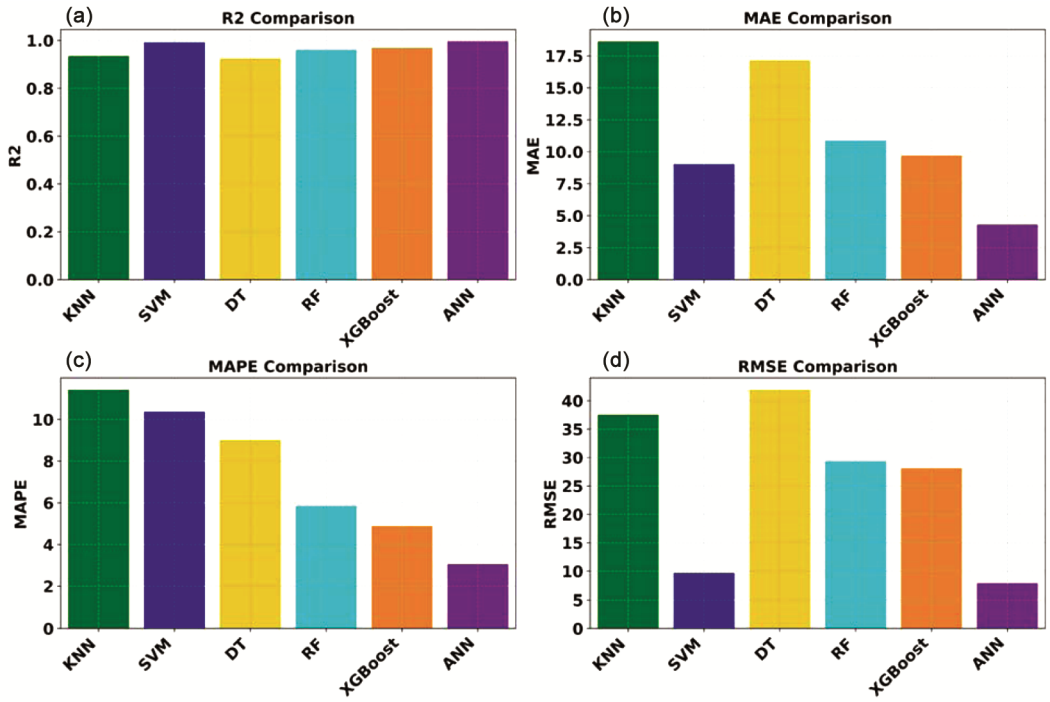

Fig. 6 — (a) R^2 Comparison of all applied ML models, (b) MAE Comparison of all applied ML models, (c) MAPE Comparison of all applied ML models, and (d) RMSE Comparison of all applied ML models (using all 7 feature).

Table 3 — Performance metrics of ML models (using top 4 features).

Model	R^2	MAE	MAPE	RMSE	Centered RMSE
K-Nearest Neighbors	0.96	16.440	11.298	31.773	31.162
Support Vector Machine	0.99	9.280	9.817	11.112	10.759
Decision Tree	0.92	18.634	12.781	42.603	42.413
RandomForestRegressor	0.96	12.876	9.452	29.183	28.954
XGBoost	0.98	12.022	9.657	21.731	21.621
ANN	0.99	8.062	8.444	10.499	10.432

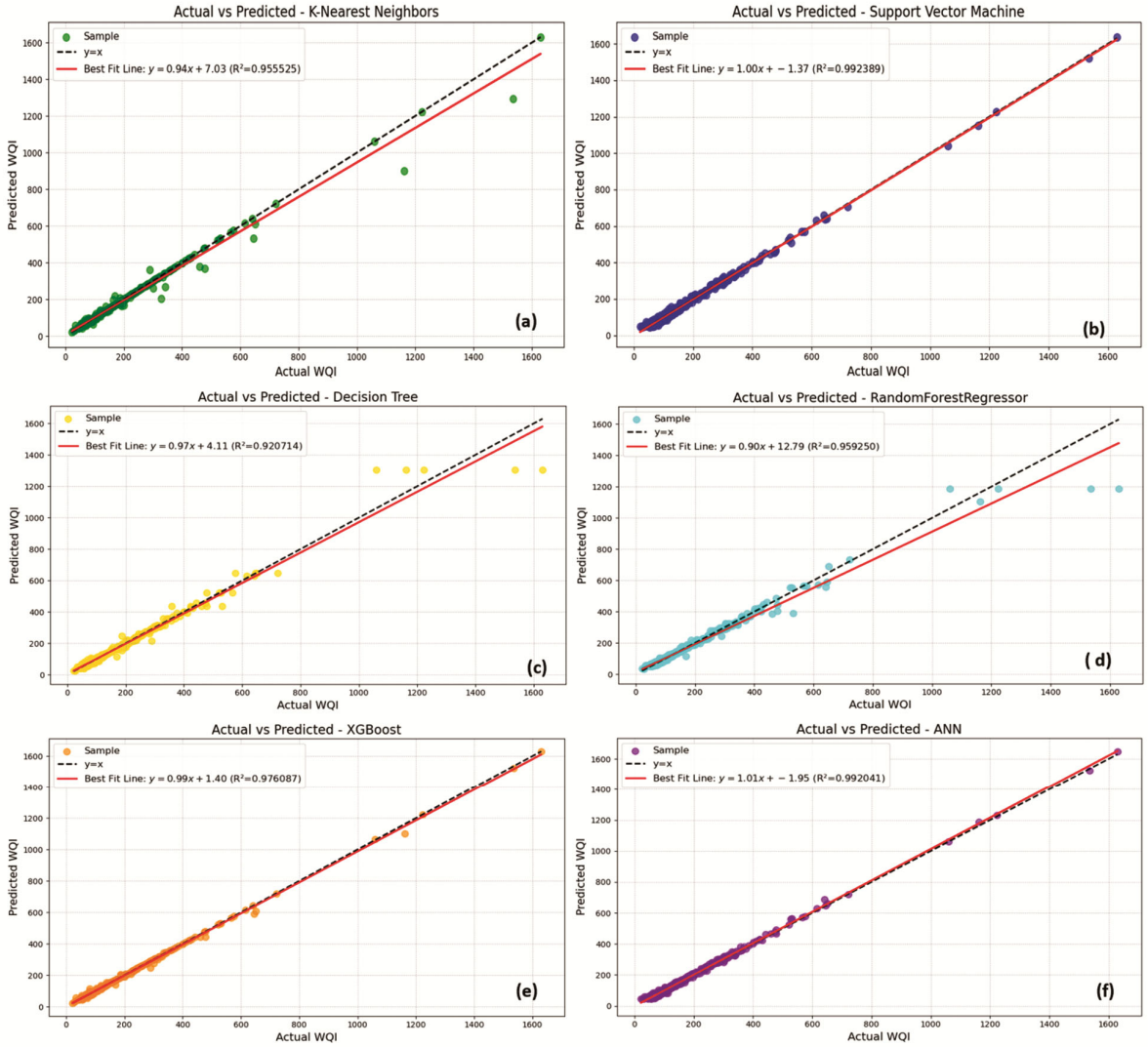

Fig. 7 — Line plot of Actual vs Predicted WQI for (a) KNN, (b) SVM, (c) DT, (d) RF, (e) XGBOOST, and (f) ANN using top 4 features.

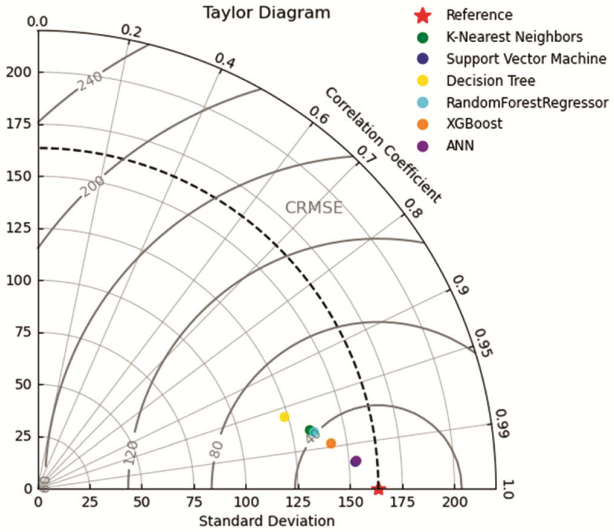

Fig. 8 — Taylor diagram (using top 4 features).

Table 4 — Models ranked on basis of distance from reference point in Taylor diagram (using top 4 features)

Rank	Model	Distance from reference
1	Support Vector Machine	17.477
2	ANN	17.541
3	XGBoost	31.848
4	RandomForestRegressor	41.332
5	K-Nearest Neighbors	43.570
6	Decision Tree	56.949

to ensure robust model evaluation. The performance metrics of various machine learning models for these features are detailed in Table 5, while Fig. 9 depicts a line plot showing actual vs. predicted WQI values for KNN, SVM, DT, RF, XGBoost, and ANN models.

3.7.1 Analysis of model accuracy for top 2 features utilizing a Taylor diagram

The Taylor diagram in Fig. 10 illustrates the performance of ML models using the top two

Table 5 — Performance metrics of ML models (using top 2 features)

Model	R^2	MAE	MAPE	RMSE	Centered RMSE
K-Nearest Neighbors	0.96	15.510	11.980	30.726	30.359
Support Vector Machine	0.99	11.821	10.733	15.527	15.060
Decision Tree	0.93	18.914	13.204	38.194	38.007
RandomForestRegressor	0.95	14.975	11.017	32.731	32.478
XGBoost	0.97	14.420	11.954	25.671	25.489
ANN	0.99	11.236	10.161	15.815	15.491

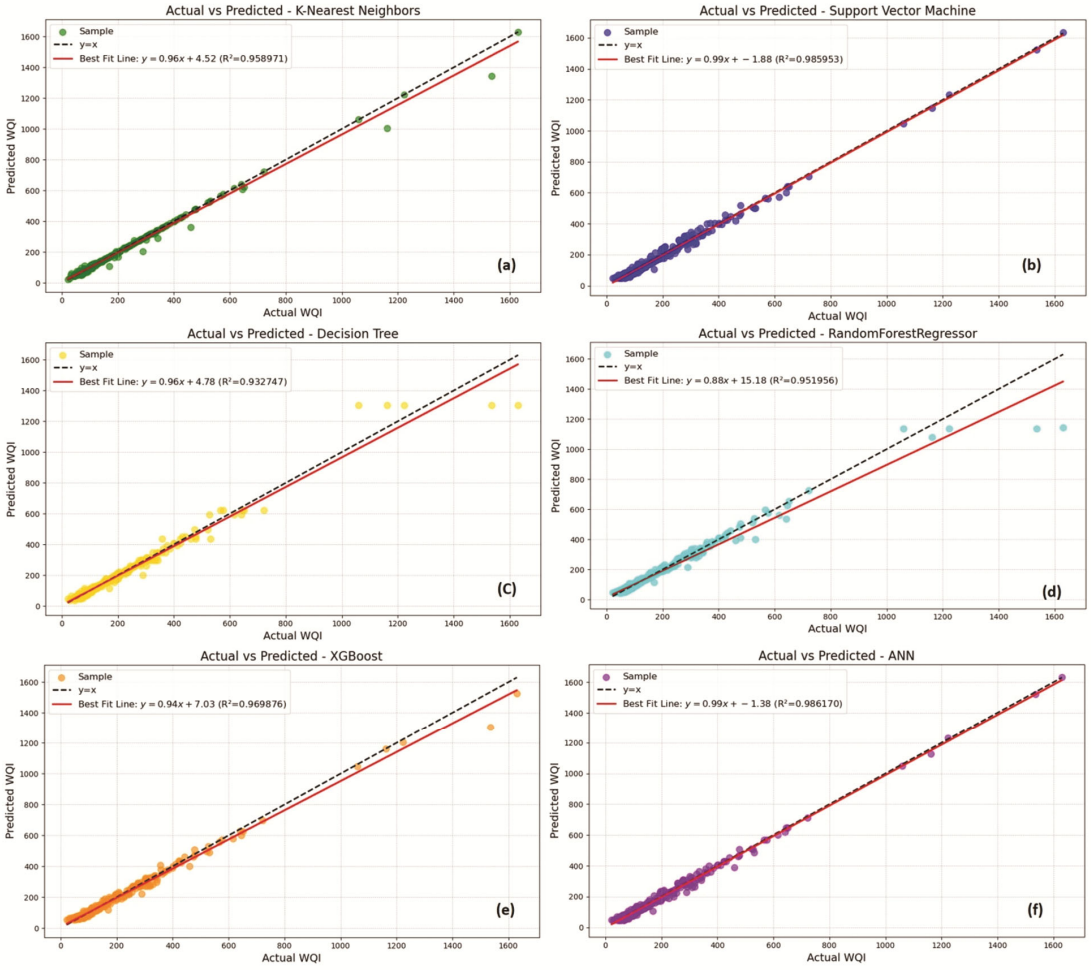

Fig. 9 — Line plot of Actual vs. Predicted WQI for (a) KNN, (b) SVM, (c) DT, (d) RF, (e) XGBOOST, and (f) ANN using top 2 features.

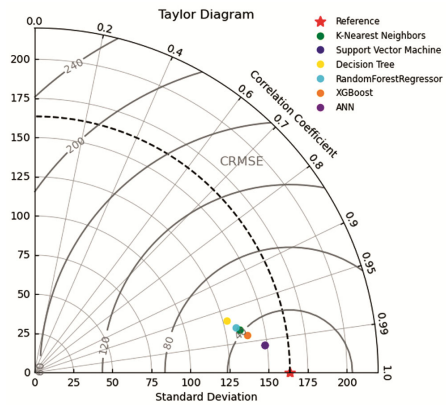

Fig. 10 — Taylor diagram (using top 2 features).

features. Table 6 ranks the models based on their distances from the reference point in the Taylor diagram, considering the top two features. Based on these rankings, the SVM model with the top two features is selected for further validation.

3.8 Analysis of model performance using top most features based on Pearson correlation coefficient to predict WQI

The top most feature used for prediction is 'K' and the target variable is WQI. The Nested K-Fold Cross Validation technique is used with k = 5. Table 7 displays the performance metrics of various machine

learning models for this feature. Consequently, Fig. 11(a-f) depicts a line plot of actual vs. predicted WQI for KNN, SVM, DT, RF, XGBoost, and ANN models using only the most significant feature.

Table 6 — Models ranked on basis of distance from reference point in Taylor diagram (using top 2 features)

Rank	Model	Distance from reference
1	Support Vector Machine	23.789
2	ANN	23.931
3	XG Boost	36.253
4	K-Nearest Neighbors	42.236
5	Random Forest Regressor	45.194
6	Decision Tree	52.376

Table 7 — Performance metrics of ML models (using K)

Model	R^2	MAE	MAPE	RMSE	Centered RMSE
K-Nearest Neighbors	0.761	39.069	28.492	64.739	64.464
Support Vector Machine	0.812	34.187	24.920	57.326	55.596
Decision Tree	0.746	39.331	27.851	69.024	68.802
Random Forest Regressor	0.784	37.460	27.064	63.456	63.174
XGBoost	0.813	35.829	26.425	59.269	58.981
ANN	0.840	34.297	27.011	52.606	52.239

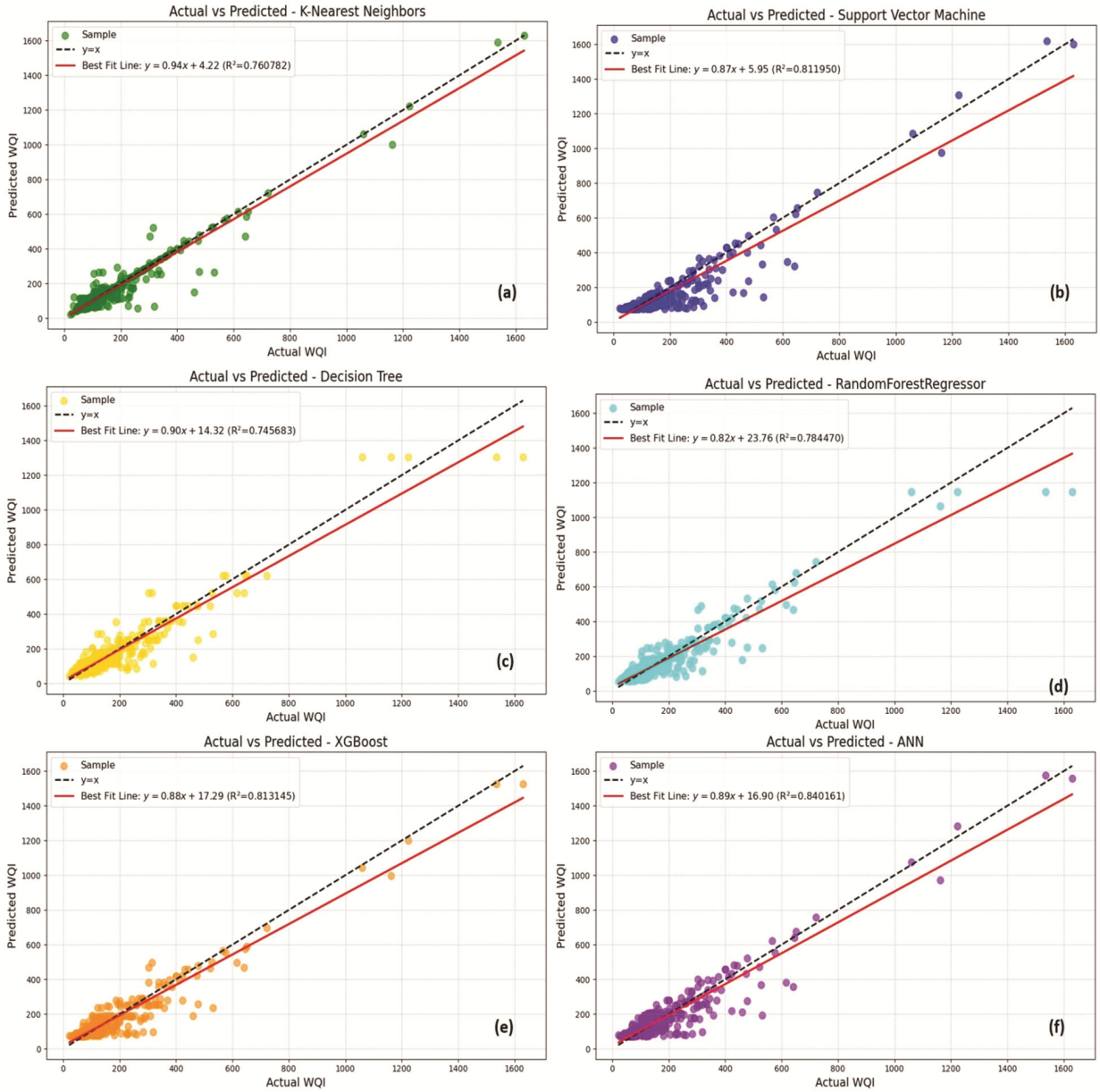

Fig. 11 — Line plot of Actual vs. Predicted WQI for (a) KNN, (b) SVM, (C) DT, (d) RF, (e) XGBOOST, and (f) ANN using top most features (K).

3.8.1 Analysis of model accuracy for top most features utilizing a Taylor diagram

The Taylor diagram in Fig. 12 illustrates the performance of machine learning models using the most significant feature. Table 8 ranks the models based on their distances from the reference point in the Taylor diagram when using this single top feature. Based on the rankings, the ANN model, utilizing only the most significant feature, is selected for further validation.

3.9 Further evaluation and validation of top-performing ML models

For each of the four cases with different input feature sets, a model was selected using the Taylor diagram for further prediction testing on an independent dataset of 70 samples. Table 9 presents the best-performing models for all four cases, while Table 10 provides the performance metrics of four. Figures 13,14,15,16 display scatter plots comparing predicted and actual WQI values for ANN (Case-1), Support Vector Machine (Case-2), Support Vector Machine (Case-3), and ANN (Case-4). Additionally, Figure 17 illustrates the Taylor diagram for further validation and performance assessment of all applied models, with Table 11 ranking the models based on their distance from the reference point in the diagram. From Tables 9 and 11, it is evident that models with a higher number of input features generally exhibited better performance. ANN (Case-1) emerged as the top-performing model; however, SVM (Case-3), which utilized only two features (Mg, K), demonstrated a very slight decrease in performance compared to ANN (Case-1), which required all seven features. Therefore, to optimize water quality monitoring, SVM (Case-3) is a practical choice, as it requires only two features, reducing both cost and time associated with laboratory testing of additional parameters.

Table 8 — Models ranked on basis of distance from reference poir Taylor diagram (using top most feature 'K')

Rank	Model	Distance from referenc
1	ANN	72.9637
2	Support Vector Machine	77.559
3	XGBoost	78.541
4	RandomForestRegressor	82.9255
5	K-Nearest Neighbors	85.7171
6	Decision Tree	88.2782

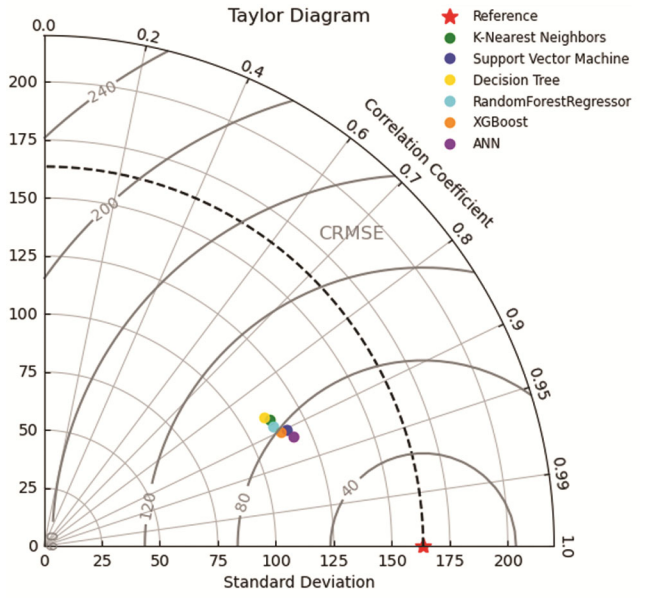

Fig. 12 — Taylor diagram (using top most feature 'K').

Table 9 — Best performing models in different cases

Case	Best Model	Input features
1	ANN	pH, EC, Mg, K, TH, Ca, Cl
2	Support Vector Machine	EC, Mg, K, TH
3	Support Vector Machine	Mg, K
4	ANN	K

Table 10 — Performance metrics of different models on a different 70 samples dataset

Model	R^2	MAE	MAPE	RMSE	Centered RMSE
ANN (Case-1)	0.999	3.746	2.550	5.520	5.373
Support Vector Machine (Case-2)	0.998	6.510	5.526	8.227	7.626
Support Vector Machine (Case-3)	0.998	6.841	5.601	9.588	9.585
ANN (Case-4)	0.932	33.369	23.210	53.087	51.699

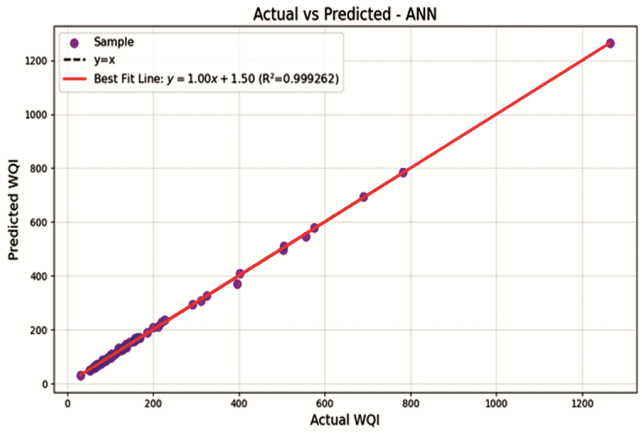

Fig. 13 — Scatter plot of Predicted vs Actual WQI of ANN (Case-1).

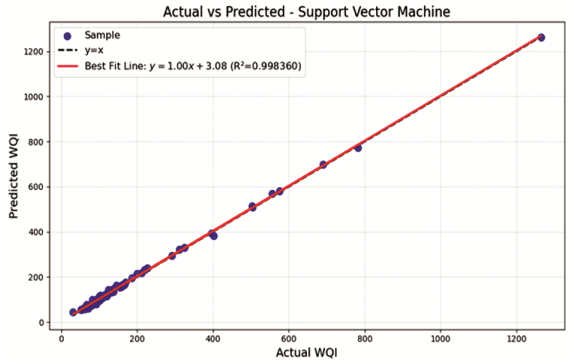

Fig. 14 — Scatter plot of Predicted vs Actual WQI of SVM (Case-2).

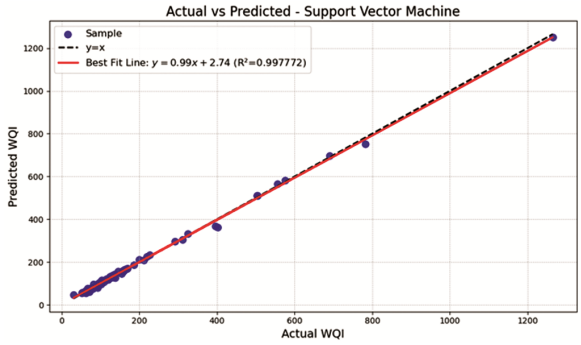

Fig. 15 — Scatter plot of Predicted vs Actual WQI of SVM (Case-3).

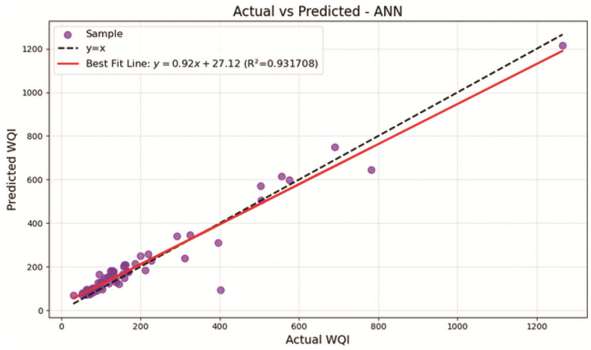

Fig. 16 — Scatter plot of Predicted vs Actual WQI of SVM (Case-4).

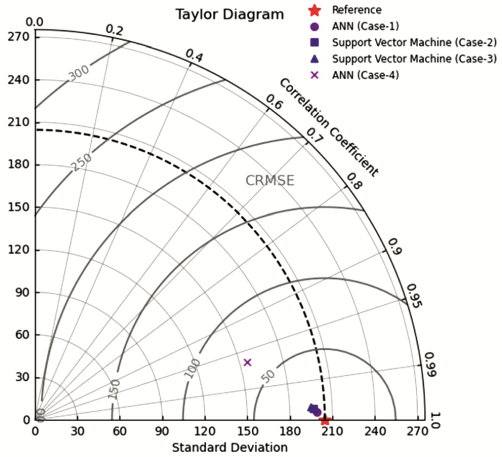

Fig. 17 —Taylor diagram for further performance check.

Table 11 — Models ranked on basis of distance from reference point in Taylor diagram

Rank	Model	Input Features	Distance
1	ANN (Case-1)	pH, EC, Mg, K, TH, Ca, Cl	7.6777
2	Support Vector Machine (Case-2)	EC, Mg, K, TH	11.1445
3	Support Vector Machine (Case-3)	Mg, K	13.4403
4	ANN (Case-4)	K	68.1224

4 Conclusion

This study provides an innovative and comprehensive evaluation of groundwater quality in the Delhi NCR, India revealing severe contamination concerns that make a significant portion of the samples unsuitable for drinking without prior treatment. The key findings of this investigation are presented below.

a Analysis of 488 groundwater samples from Delhi NCR revealed high concentrations of EC, Cl, Mg, and TH, with many samples exceeding permissible limits for drinking water set by Indian Standard (IS 10500:2012) and World Health Organization (WHO, 2017). These contaminants pose significant risks to human health, agricultural viability, and environmental stability.

b Six machine learning models (KNN, SVM, DT, RF, XGBoost, ANN) were employed to predict the Water Quality Index (WQI). Among them, Artificial Neural Networks (ANN) had the greatest predicted accuracy utilizing all seven parameters. The Support Vector Machine (SVM) demonstrated robust efficacy with only 2–4 input parameters, providing a cost-efficient solution for water quality prediction.

c The study showed four parameter-reduction scenarios using Pearson correlation to identify minimal yet effective feature sets. SVM demonstrated reliability solely with Mg and K, while ANN showed robust performance exclusively with K, though with some limitations.

d Model validation utilizing an independent dataset of 70 samples, accompanied by Taylor diagram analysis, revealed ANN (with all features) as the most accurate model, followed by SVM (with reduced features).

e Geospatial Analysis revealed spatial trends and contamination clusters around key urban zones, particularly near Bhalswa landfill. This offers practical information for focused groundwater monitoring and remediation solutions.

f This research directly supports Sustainable Development Goals by promoting efficient, data-driven groundwater quality assessment. The integration of machine learning techniques reduces the reliance on extensive laboratory testing, enabling cost-effective and timely monitoring. This promotes equitable access to clean and affordable drinking water and enhances sustainable water management techniques in rapidly urbanizing areas such as DelhiNCR, India.

g The study demonstrates a novel integration of machine learning and geospatial analytics for innovative, scalable, and sustainable water quality monitoring. This framework can reduce reliance on intensive laboratory testing, lower costs, and improve early warning systems for groundwater contamination.

Acknowledgements

The authors are thankful to Delhi Technological University, Delhi, India, for providing the computing resources, data collection, and Laboratory support necessary for this work.

References

1 Velis M, Conti K I & Biermann F, *Sustain Sci*, 12 (2017) 1007–1017.
2 Raheja H, Goel A & Pal M, *Water Pract Technol*, 17 (2021) 336–351.
3 Mohamed I, Othman F, Ibrahim A I N, Alaa-Eldin M E & Yunus R M, *Environ Monit Assess*, 187 (2014) 1–17.
4 Bhardwaj D & Verma N, *Int J Adv Res Comput Sci*, 8 (2017) 2496–2498.
5 Malek N H A, Wan Yaacob W F, Md Nasir S A & Shaadan N, *Water*, 14 (2022) 1067.
6 Molajou A, Afshar A, Khosravi M, Soleimanian E, Vahabzadeh M & Variani H A, *Environ Sci Pollut Res*, 30 (2021) 107487–107497.
7 Tian H, Huang N, Niu Z, Qin Y, Pei J & Wang J, *Remote Sens*, 11 (2019) 820.
8 Vasanthavigar M, Srinivasamoorthy K, Vijayaragavan K, Rajiv Ganthi R, Chidambaram S, Anandhan P, Manivannan R & Vasudevan S, *Environ Monit Assess*, 171 (2010) 595–609.
9 Zahedi S, *Ecol Indic*, 83 (2017) 368–379.
10 Kumari M & Rai S C, *J Geol Soc India*, 95 (2020) 159–168.
11 Maghrebi M, Noori R, Partani S, Araghi A, Barati R, Farnoush H & Torabi Haghighi A, *Earth Space Sci*, 8 (2021).
12 Haghiabi A H, Nasrolahi A H & Parsaie A, *Water Qual Res J*, 53 (2018) 3–13.
13 Lu H & Ma X, *Chemosphere*, 249 (2020) 126169.
14 Nayan A-A, Kibria M G, Rahman Md O & Saha J, *Adv Inf Commun Technol*, 2020 (2020) 219–224.
15 Lumb A, Sharma T C, Bibeault J-F & Klawunn P, *Water Qual Expo Health*, 3 (2011) 203–216.
16 Dao V, Urban W & Hazra S B, *Groundwater Sustain Dev*, 11 (2020) 100457.
17 Mohammadpour R, Shaharuddin S, Chang C K, Zakaria N A, Ghani A A & Chan N W, *Environ Sci Pollut Res*, 22 (2014) 6208–6219.
18 Singh K P, Basant N & Gupta S, *Anal Chim Acta*, 703 (2011) 152–162.
19 Safavi H R & Esmikhani M, *Water Resour Manag*, 27 (2013) 2623–2644.
20 Abba S I, Pham Q B, Saini G, Linh N T T, Ahmed A N, Mohajane M, Khaledian M, Abdulkadir R A & Bach Q-V, *Environ Sci Pollut Res*, 27 (2020) 41524–41539.
21 El Bilali A & Taleb A, *Phys Chem Earth Parts A/B/C*, 136 (2024) 103794.
22 Central Ground Water Board, *Ground Water Quality Data Report, Delhi NCR*, (2021).
23 Hyarat T, Al Kuisi M & Saffarini G, *Water Pract Technol*, 17 (2022) 1582–1602.
24 Rana A, Singh Rawat A, Bijalwan A & Bahuguna H, *Res Intell Comput Eng*, 2018 (2018) 1–6.
25 Sutadian A D, Muttil N, Yilmaz A G & Perera B J C, *Environ Monit Assess*, 188 (2015) 1.
26 Balan I, Madan Kumar P & Shivakumar M, *Chron Young Sci*, 3 (2012) 146.
27 Krishan A, Mishra R K & Khursheed A, *Urban Water J*, 19 (2022) 520–530.
28 Shaveta, *Int J Sci Res Arch*, 9 (2023) 281–285.
29 Sarker I H, *SN Comput Sci*, 2 (2021).
30 Zhu M, Wang J, Yang X, Zhang Y, Zhang L, Ren H, Wu B & Ye L, *Eco-Environment Health*, 1 (2022) 107–116.
31 Loh W, *Int Stat Rev*, 82 (2014) 329–348.
32 Zacharis N Z, *Int J Intell Syst Appl*, 10 (2018) 1–9.
33 Klusowski J M, *arXiv preprint arXiv:1906.10086*, (2019).
34 Bagheri S, Taridashti S, Farahani H, Watson P & Rezvani E, *Front Psychiatry*, 14 (2023).
35 Madhiarasan M & Louzazni M, *J Electr Comput Eng*, 2022 (2022) 1–23.
36 Chan K Y, Abu-Salih B, Qaddoura R, Al-Zoubi A M, Palade V, Pham D-S, Ser J D & Muhammad K, *Neurocomputing*, 545 (2023) 126327.
37 Opitz D & Maclin R, *J Artif Intell Res*, 11 (1999) 169–198.
38 Sagi O & Rokach L, *WIREs Data Min Knowl Discov*, 8 (2018).
39 Shu B, Liu Y, Wang C, Zhang H, Amani-Beni M & Zhang R, *Ecol Indic*, 166 (2024) 112554.
40 Schonlau M & Zou R Y, *Stata J Promot Commun Stat Stata*, 20 (2020) 3–29.
41 Alomari A H, Al-Mistarehi B W, Alnaasan T K & Obeidat M S, *Appl Sci*, 13 (2023) 5113.
42 Asamoah E, Heuvelink G B M, Chairi I, Bindraban P S & Logah V, *Heliyon*, 10 (2024) e37065.
43 Tyralis H, Papacharalampous G & Langousis A, *Water*, 11 (2019) 910.
44 Zhang W, Liu D & Cao K, *Case Stud Constr Mater*, 21 (2024) e03416.
45 Chen T & Guestrin C, *Knowl Discov Data Min*, 22 (2016) 785–794.
46 Niazkar M, Menapace A, Brentan B, Piraei R, Jimenez D, Dhawan P & Righetti M, *Environ Model Softw*, 174 (2024) 105971.
47 Cover T & Hart P, *IEEE Trans Inf Theory*, 13 (1967) 21–27.
48 Ebrahimi M & Basiri A, *Knowl-Based Syst*, 301 (2024) 112357.

49 Goyal R, Chandra P & Singh Y, *IERI Procedia*, 6 (2014) 15–21.
50 Liu H, He S & Peng J, *Heliyon*, 10 (2024) e33781.
51 Figueiredo Filho D B, Silva Júnior J A & Rocha E C, *Leviathan (São Paulo)*, 3 (2011) 60.
52 Robeson S M & Willmott C J, *PLOS ONE*, 18 (2023) e0279774.
53 Kim S & Kim H, *Int J Forecast*, 32 (2016) 669–679.
54 Javadi F, Qaderi K, Ahmadi M M, Rahimpour M, Madadi M R & Mahdavi-Meymand A, *Sci Rep*, 12 (2022).
55 Taylor K E, *J Geophys Res Atmos*, 106 (2001) 7183–7192.
56 Huang R, Hanif M F, Siddiqui M K & Hanif M F, *Comput Mater Sci*, 240 (2024) 112994.
57 Ling Y, Gong L, Ni G, Zhang X, Li Z & Li M, *J Mol Liq*, 411 (2024) 125734.
58 Zhao T, Zheng Y & Wu Z, *Comput Chem Eng*, 169 (2023) 108074.
59 Chandrashekar G & Sahin F, *Comput Electr Eng*, 40 (2014) 16–28.
60 Rendall R, Castillo I, Schmidt A, Chin S-T, Chiang L H & Reis M, *Comput Chem Eng*, 121 (2019) 99–110.
61 Brown R M, McClelland N I, Deininger R A & O'Connor M F, *Indic Environ Qual*, (1972) 173–182.
62 Bharadwaj S, Gupta A K & Sahu A K, *Adv Constr Manag Lect Notes Civ Eng*, 601 (2025).