MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Integrating Machine Learning Techniques to Assess and Predict Water Contamination in the Gomti River Basin

Nidhi Singh[1], Smita Tung[2]
[1]Research Scholar, Civil Engineering Department, GLA University, Mathura, India, nidhisingh.abc@gmail.com (Corresponding Author)
[2]Assistant Professor, Civil Engineering Department, GLA University, Mathura, India, smita.tung@gla.ac.in

Abstract: The increasing contamination of river systems due to rapid urbanization, industrial discharge, and agricultural runoff poses a serious threat to environmental and public health. Traditional methods often fall short in capturing the complex, non-linear interactions among multiple water quality parameters. To address this gap, the present study applies advanced machine learning (ML) models to assess and predict the water quality index (WQI) in the Gomti river basin during the pre-monsoon season, a period with minimal dilution effects. A total of 100 samples were collected from five strategic sites, and 18 physicochemical and heavy metal parameters were analyzed. The study compares the performance of multiple linear regression (MLR), artificial neural networks (ANN), support vector regression (SVR), random forest regression (RFR), and decision tree regression (DTR). Results show that ANN achieved the highest predictive accuracy (92.52% of predictions within ±20% of actual WQI), followed by RFR (88%), while MLR and DTR showed limited performance. Feature importance and sensitivity analysis identified electrical conductivity (EC) dissolved oxygen (DO), total dissolved solids (TDS), and sulfate (SO_4) as the most influential predictors of WQI. This study demonstrates the potential of ML-based models for accurate water quality prediction and supports data-driven strategies for sustainable water resource management. Future work should incorporate seasonal variations, real-time sensor data, and hybrid modeling frameworks to enhance predictive reliability and support early-warning systems.

Keywords: Water quality index (WQI); Gomti river basin; machine learning models; pre-monsoon contamination; environmental stressors.

1. Introduction

Water functions as a critical resource for sustaining ecological systems, agricultural productivity, industrial operations, and public health (Das et al. 2024). The deterioration of water quality due to escalating anthropogenic pressures—such as agricultural runoff, industrial effluents, and rapid urbanization—has emerged as a global concern, compromising the sustainability of freshwater ecosystems (Das et al., 2025; Ayvaz, 2010; Jiang et al., 2018; Zheng et al., 2024). These issues are closely aligned with the United Nations Sustainable Development Goal 6 (SDG 6), which advocates for universal access to clean water and integrated water resource management (WHO 2008).Effective water quality monitoring is essential to detect environmental stressors, support pollution control strategies, and preserve aquatic health (Surya Prakash 2014). However, predicting surface water quality remains complex due to the nonlinear interactions among diverse physicochemical, hydrological, and climatic factors (Khan et al. 2022). Conventional statistical approaches, typically grounded in linear assumptions, often lack the flexibility to model the multifactorial and dynamic nature of water quality systems (Singh et al. 2009; Singh and Datta 2007).Recent advances in machine learning (ML) have introduced adaptive, data-driven techniques capable of modeling nonlinear, high-dimensional relationships and improving predictive accuracy (Ahmed, Mumtaz, and Zaidi 2021; Das et al., 2025; Das et al. 2024). Models such as artificial neural networks (ANN), support vector regression (SVR), random forest regression (RFR), decision tree regression (DTR), and multiple linear regression (MLR) have demonstrated considerable success in environmental modeling and classification tasks (Kushwah et al. 2023; Saqib et al. 2023). These models offer capabilities to integrate multivariate datasets, uncover hidden patterns, and support decision-making under uncertainty, particularly in the domain of surface water quality assessment (Das et al., 2025; Das et al., 2024).Despite these advancements, studies applying ML-based predictive frameworks in the Gomti River Basin remain scarce. Located in northern India, the Gomti Basin is ecologically significant and supports a densely populated region, yet faces mounting challenges related to untreated wastewater discharge, agricultural intensification, and land use change. Most existing investigations within the basin have focused on empirical or descriptive water quality evaluations, often neglecting predictive modeling, comparative algorithm testing, or sensitivity analysis (Ahmed et al. 2021; Khan et al. 2022; Surya Prakash 2014).

Recent literature has introduced innovative methods such as fuzzy multi-criteria decision-making, weighted hesitant fuzzy soft sets, and neutrosophic models for evaluating urban river water quality (Das et al., 2025; Das et al., 2024; Das & Granados, 2024). Although methodologically diverse, these approaches often lack real-time predictive capabilities and face challenges in model interpretability, scalability, and integration with high-frequency field datasets. Moreover, few studies have explicitly linked predictive analytics with region-specific management strategies or policy frameworks (Kanti et al. 2024; Mohinuddin et al. 2023; Singh et al. 2009; Singh and Datta 2007).

This study addresses these gaps by applying and comparing five machine learning models—MLR, ANN, SVR, RFR, and DTR—for prediction of water quality index (WQI) in the Gomti River Basin, using pre-monsoon primary data. Model performance is assessed using multiple accuracy metrics, and a sensitivity analysis identifies the most influential water quality parameters. The study also evaluates how ML-based findings can inform policy recommendations under varying pollution scenarios. The key objectives are as follows; a) To identify the dominant water quality parameters influencing WQI at selected sites in the Gomti River Basin; b) To compare the predictive performance of MLR, ANN, SVR, RFR, and DTR models for WQI estimation; c) To perform sensitivity analysis for model interpretability and parameter prioritization; and, d) To provide data-driven recommendations for regional water quality governance and policy development. This research offers a replicable framework for ML-based water quality modeling and supports alignment with national water quality monitoring efforts and global sustainability goals. The subsequent sections present a detailed literature review, description of the study area, sampling procedures, modeling methodologies, results, analysis, limitations, and future implications.

2. Study Area

The Gomti river, a significant tributary of the Ganges, originates from Gomat Taal in Pilibhit, Uttarakhand, and flows for approximately 940 kilometers through the Indian state of Uttar Pradesh. It traverses fertile alluvial plains that support more than 15 million people, making the basin critically important for agriculture, urban development, and ecological sustainability. The river's role in meeting the water demands of both rural and urban populations underlines its socio-economic and environmental importance.Climatically, the region experiences a humid subtropical climate, with three distinct hydrological seasons: pre-monsoon (hot and dry), monsoon (wet and humid), and post-monsoon (cool and dry). The pre-monsoon period, characterized by minimal rainfall and reduced river discharge, presents an ideal condition for monitoring pollution concentrations due to the absence of monsoonal dilution effects (Pimparkar et al. 2023). The average annual rainfall in the basin is about 1,000 mm, concentrated largely in the monsoon season. Concurrently, intensive groundwater extraction—driven by agricultural, domestic, and industrial demands—further stresses the river system and impacts both surface and subsurface hydrology.The Gomti river basin faces increasing environmental and anthropogenic pressures. Rapid urbanization, especially in urban centers such as Lucknow, contributes significant volumes of untreated domestic and industrial wastewater directly into the river. In parallel, agricultural runoff, enriched with fertilizers and pesticides, promotes eutrophication and disrupts aquatic ecosystems. These challenges are compounded by deforestation and unplanned land-use changes, which accelerate sedimentation and reduce the river's ecological resilience (Bose, Mele, and Silberschmidt 2024; Saqib et al. 2023; Sohn 2021)

To capture spatial heterogeneity and variations in water quality across the basin, five strategically selected sampling sites were identified, representing diverse upstream, midstream, and downstream conditions. At each site, 20 water samples were collected during the pre-monsoon season of 2024. This sampling strategy allowed for a detailed assessment of physicochemical and heavy metal parameters across varying environmental gradients and anthropogenic influence zones.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Figure 1 illustrates the Gomti river basin, marking the five sampling locations. The spatial mapping provides essential context for understanding the distribution of contaminants, enabling a site-specific interpretation of the river's water quality status.

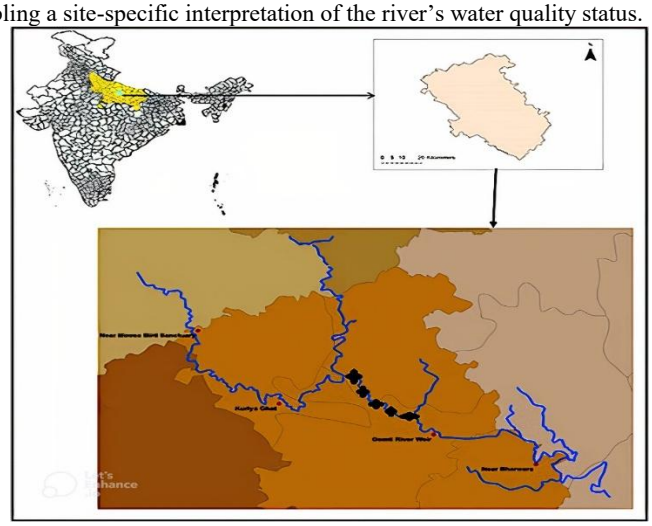

Figure 1. Map of the Gomti River Basin Showing Sample Collection Points

3. Methodology

3.1 Sample Collection, Laboratory Testing, and Data Preprocessing

Water quality samples were collected from five strategically selected sites across the Gomti river basin to capture spatial variability in hydro-environmental conditions and potential contamination. The sites were chosen based on their proximity to urban wastewater discharge areas, zones of intensive agricultural runoff, and relatively undisturbed upstream locations. This sampling strategy ensured diverse representation of land-use influences and pollution sources, allowing for a comprehensive assessment of water quality conditions across the basin.

A total of 100 surface water samples, 20 per site, were collected over a three-week period during the pre-monsoon season in May 2024. This period was deliberately selected due to minimal rainfall and low river discharge, which reduces dilution and enhances the detectability of both point and non-point source pollutants (Haq, Jilani, and Prabu 2022; Sidek et al. 2024). Standard grab sampling procedures, as outlined by WHO (2008), were followed. Samples were collected in acid-washed, pre-cleaned polyethylene bottles, stored in insulated carriers with ice packs, and transported to the laboratory within six hours to prevent any alteration in physicochemical properties. In-field measurements of pH and electrical conductivity (EC) were recorded using calibrated portable meters.In the laboratory, a comprehensive suite of water quality parameters was analyzed, including physicochemical indicators such as EC, pH, total dissolved solids (TDS), total hardness, dissolved oxygen (DO), and biochemical oxygen demand (BOD); major anions and cations including fluoride (F^-), nitrate (NO_3^-), sulfate (NO_4^{2-}), chloride (Cl^-), calcium (Ca^{2+}), and magnesium (Mg^{2+}); and trace heavy metals such as cadmium (Cd), lead (Pb), zinc (Zn), chromium (Cr), iron (Fe), manganese (Mn), and copper (Cu). These parameters were selected due to their ecological significance and known health implications (Khan et al. 2021; Khan, Saxena, and Shukla 2020).Analytical techniques were chosen based on accuracy, sensitivity, and compliance with national and international standards (Das et al. 2019). Spectrophotometric, titrimetric, and electrometric methods were used for general water quality parameters. Calcium and magnesium were quantified using flame atomic absorption spectroscopy (AAS), while trace heavy metals were measured using inductively coupled plasma optical emission spectroscopy (ICP-OES) for its enhanced detection capabilities. A full summary of the analyzed parameters, their abbreviations, units, and corresponding methods is presented in Table 1.

To ensure data validity and reproducibility, a robust quality assurance and quality control (QA/QC) protocol was implemented. Instruments were calibrated daily using certified reference standards. Analytical batches included method blanks, duplicates, and spiked recovery samples to monitor precision and accuracy. Acceptable recovery rates of 90–110% and relative standard deviations below 10% confirmed the reliability and integrity of the dataset for modeling purposes (Kushwah et al. 2023; Mohinuddin et al. 2023).

Before initiating machine learning analysis, the dataset underwent thorough preprocessing. Outliers were identified and removed using Z-score analysis ($|Z| > 3$), and missing values, which were minimal, were imputed using the mean. To ensure consistency and support algorithm convergence, all variables were normalized using min–max scaling (Das et al. 2019). The processed dataset was reviewed for completeness and internal consistency, confirming its suitability for both statistical evaluation and predictive modeling.

Table 1. Summary of Input and Output Parameters with Analytical Methods

Type	Parameter	Abbreviation	Units	Analytical Method
Input	Electrical Conductivity	EC	µS cm^{-1}	Electrometric
Input	pH-Value	pH	pH unit	pH-meter
Input	Total Dissolved Solids	TDS	mg l^{-1}	Gravimetric
Input	Total Hardness	TH	$CaCO_3$ mg l^{-1}	Titrimetric
Input	Dissolved Oxygen	DO	mg l^{-1}	Winkler azide method
Input	Biochemical oxygen demand	BOD	mg l^{-1}	Winkler azide method
Input	Chloride	Cl	mg l^{-1}	Spectrophotometric
Input	Fluoride	F	mg l^{-1}	Spectrophotometric
Input	Sulphate	SO_4	mg l^{-1}	Spectrophotometric
Input	Calcium	Ca	mg l^{-1}	Flame Atomic Absorption Spectroscopy (AAS)
Input	Magnesium	Mg	mg l^{-1}	Flame Atomic Absorption Spectroscopy (AAS)
Input	Nitrate nitrogen	NO_3	mg l^{-1}	Spectrophotometric
Input	Cadmium	Cd	mg l^{-1}	Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES)
Input	Chromium	Cr	mg l^{-1}	ICP-OES
Input	Iron	Fe	mg l^{-1}	ICP-OES
Input	Lead	Pb	mg l^{-1}	ICP-OES
Input	Copper	Cu	mg l^{-1}	ICP-OES
Input	Manganese	Mn	mg l^{-1}	ICP-OES
Input	Zinc	Zn	mg l^{-1}	ICP-OES
Output	Water Quality Index	WQI	-	Calculated using formulas

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

3.2 Water Quality Index (WQI) Calculation

The WQI was employed as a composite metric to assess the suitability of surface water for general use by aggregating multiple physicochemical indicators into a single standardized value. This index facilitates the interpretation of complex water quality data and enables comparative assessment across sites with varying contamination profiles (Zheng et al. 2024). To calculate the WQI, each parameter was first normalized to a quality rating scale (0–100) using the formula shown in Eq. (1):

$$q_i = \frac{V_{actual} - V_{ideal}}{V_{standard} - V_{ideal}} \times 100 \quad (1)$$

where V_{actual} is the observed value, V_{ideal} is the ideal value (typically zero for pollutants), and $V_{standard}$ is the permissible limit as per WHO and BIS standards. Parameters were assigned weights (W_i) based on their relative importance to water quality and ecological health. The WQI was then computed using the Eq. (2):

$$WQI = \frac{\sum q_i . W_i}{\sum W_i} \quad (2)$$

The WQI values were subsequently categorized into five qualitative classes: **excellent (0–25), good (26–50), moderate (51–75), poor (76–100),** and **very poor (>100),** enabling a standardized interpretation of water quality status (Nandi et al. 2024; Saalidong et al. 2022). Table 2 presents descriptive statistics for 19 input parameters and the WQI. EC values ranged from 220 to 613 µS/cm, with a mean of 406.78 and a standard deviation of 80.34. Elevated EC at some sites suggests high ionic content, possibly due to domestic sewage or agricultural runoff. A slight positive skew (0.51) and moderate kurtosis (0.14) indicate a fairly symmetric distribution with few outliers. EC is a major indicator of salinity and correlates strongly with TDS. The pH ranged from 7.43 to 8.77, with a mean of 8.24 and low variability (SD = 0.30). Slight alkalinity may be attributed to biological activity, industrial effluents, or carbonate buffering. A negative skewness (-0.86) indicates clustering toward higher values. All values lie within permissible drinking water limits, suggesting moderate buffering capacity of the river.

TDS values ranged from 154.53 to 372.82 mg/L, averaging 262.05 mg/L. Variability is moderate (SD = 47.73), and low kurtosis (–0.71) suggests a flatter distribution. TDS reflects the total inorganic salts and is influenced by leaching, urban discharge, and evaporative concentration. Higher values were observed near urban and downstream regions. Hardness varied widely from 52 to 280 mg/L (mean = 174.02, SD = 51.83). This reflects the presence of calcium and magnesium ions, likely from natural geologic sources or runoff. Slight negative skew (-0.36) and low kurtosis (–0.45) suggest a relatively symmetric distribution with no extreme values.DO values ranged between 0.52 and 13.5 mg/L (mean = 6.09). Low DO at some locations indicates oxygen depletion due to microbial degradation of organic matter. The distribution is symmetric (skew = 0.13), with moderate spread (SD = 2.43). DO is a critical parameter for aquatic life and inversely related to BOD. BOD varied from 1.10 to 25.54 mg/L, with a mean of 8.38 and SD of 5.90, indicating substantial variability. High BOD levels in urban locations point to organic waste contamination. Positive skew (0.81) and low kurtosis (–0.28) suggest the presence of outliers due to intense local pollution.

Cl ranged from 0.95 to 25.44 mg/L, averaging 7.75 mg/L. This wide variation (SD = 5.45) likely stems from domestic sewage and agricultural inputs. Positive skew (1.10) and moderate kurtosis (0.52) reflect outlier-influenced distribution, particularly at urban discharge points. Fluoride ranged from 0.10 to 0.93 mg/L (mean = 0.46). Variability is moderate (SD = 0.19). The data shows a roughly symmetric spread (skew = –0.01), with levels mostly below the permissible limit of 1 mg/L. Natural geogenic sources may explain the spatial distribution.

Sulfate concentrations varied from 2.9 to 25.81 mg/L (mean = 12.31), showing relatively high standard deviation (6.20). This anion is associated with industrial discharges and detergents. Negative kurtosis (–1.07) indicates a flatter, wide-ranging distribution across sampling sites. Ca ranged from 5.10 to 69.19 mg/L (mean = 39.34), with notable variation (SD = 12.72). The flat kurtosis (–0.38) and minimal skew (0.01) suggest a uniform distribution. Calcium originates from natural sources and contributes to overall hardness. Magnesium concentrations ranged from 1.50 to 40.60 mg/L (mean = 18.60, SD = 9.15). Slight positive skew (0.10) and near-normal kurtosis (–0.26) indicate a balanced distribution. The variation may reflect differences in geological formations and input from fertilizers. Nitrate ranged from 0.02 to 1.85 mg/L (mean = 0.42). The sharp positive skew (1.71) and high kurtosis (2.55) indicate several high-value outliers, likely from fertilizer runoff. The low mean suggests generally safe levels, with concerns at specific hotspots.

Cd levels were extremely low (mean = 0.00), with no variation detected (SD = 0.00). Though statistically negligible here, its inclusion is important due to toxicity risks. The absence of variation may be due to detection limits or low regional presence. Cr ranged from 0.00 to 0.04 mg/L, with a mean of 0.01. The skew (1.70) and high kurtosis (3.25) indicate a few elevated values, suggesting localized industrial contamination. Overall levels remain below critical thresholds. Fe concentrations ranged from 0.00 to 16.28 mg/L (mean = 3.26), showing high variability (SD = 3.29) and strong positive skew (1.36). This suggests industrial or geologic influence, especially near mining or construction zones. Pb ranged from 0.00 to 0.08 mg/L (mean = 0.03), with skewness of 0.56. This moderate asymmetry, along with low kurtosis, indicates some urban sites may contribute to lead presence through runoff from plumbing or vehicular residues. Cu levels varied between 0.00 and 0.21 mg/L (mean = 0.03). While positively skewed (2.31), concentrations are mostly low. Spikes may be due to corrosion from pipes or specific industrial effluents.

Mn ranged from 0.00 to 0.67 mg/L (mean = 0.16), showing substantial variability (SD = 0.14). The distribution is moderately skewed (1.02) and leptokurtic (0.83), suggesting some anomalous readings, possibly from industrial waste. Zn ranged from 0.00 to 0.56 mg/L, averaging 0.10. Skewness (1.72) and kurtosis (4.18) indicate significant outliers. Zinc may originate from galvanization industries or leaching from infrastructure. WQI values spanned 4.04 to 155.66, with a mean of 38.46. The wide spread (SD = 28.28) and skew (1.49) suggest site-specific pollution hotspots. WQI encapsulates multi-parameter influence, making it ideal for evaluating overall water health.

Table 2. Descriptive Statistics of Input WQ Parameters

Parameters	EC	pH	TDS	Hardness	DO	BOD	Cl	F	SO_4	Ca	Mg	NO3	Cd	Cr	Fe	Pb	Cu	Mn	Zn	WQI
Mean	406.78	8.24	262.05	174.02	6.09	8.38	7.75	0.46	12.31	39.34	18.60	0.42	0.00	0.01	3.26	0.03	0.03	0.16	0.10	38.46
SE	8.03	0.03	4.77	5.18	0.24	0.59	0.54	0.02	0.62	1.27	0.91	0.04	0.00	0.00	0.33	0.00	0.00	0.01	0.01	2.83
SD	80.34	0.30	47.73	51.83	2.43	5.90	5.45	0.19	6.20	12.72	9.15	0.43	0.00	0.01	3.29	0.02	0.04	0.14	0.11	28.28
SV	6454.87	0.09	2277.71	2686.63	5.89	34.86	29.69	0.04	38.39	161.84	83.72	0.18	0.00	0.00	10.82	0.00	0.00	0.02	0.01	799.72
Kurtosis	0.14	0.40	-0.71	-0.45	-0.24	-0.28	0.52	-0.75	-1.07	-0.38	-0.26	2.55	0.26	3.25	2.37	-0.87	5.66	0.83	4.18	3.04
Skewness	0.51	-0.86	0.17	-0.36	0.13	0.81	1.10	-0.01	0.42	0.01	0.10	1.71	1.03	1.70	1.36	0.56	2.31	1.02	1.72	1.49
Minimum	220.00	7.43	154.53	52.00	0.52	1.10	0.95	0.10	2.90	5.10	1.50	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.04
Maximum	613.33	8.77	372.82	280.00	13.50	25.54	25.44	0.93	25.81	69.19	40.60	1.85	0.00	0.04	16.28	0.08	0.21	0.67	0.56	155.66

The relationships among water quality parameters were explored using a correlation matrix, presented in Table 3. Parameters such as DO and BOD showed a significant inverse correlation (-0.55), consistent with their ecological interaction. Understanding these correlations is crucial for identifying key drivers of WQI variability and optimizing machine learning models for predictive accuracy.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Table 3. Correlation Matrix of Input Parameters

	EC	pH	TDS	Hardness	DO	BOD	Cl	F	SO_4	Ca	Mg	NO_3	Cd	Cr	Fe	Pb	Cu	Mn	Zn	WQI
EC	1.00																			
pH	-0.22	1.00																		
TDS	0.32	-0.30	1.00																	
Hardness	0.24	-0.19	-0.09	1.00																
DO	-0.29	0.44	-0.29	-0.14	1.00															
BOD	0.35	-0.46	0.23	0.15	-0.55	1.00														
Cl	0.20	-0.28	0.18	0.21	-0.24	0.32	1.00													
F	0.10	-0.15	0.24	0.18	-0.31	0.27	0.28	1.00												
SO_4	0.25	-0.06	0.14	0.06	-0.16	0.24	0.49	0.08	1.00											
Ca	-0.04	0.07	0.00	-0.03	0.11	0.19	0.10	0.01	0.05	1.00										
Mg	-0.04	-0.14	0.09	0.00	-0.11	0.11	-0.05	0.11	-0.04	-0.13	1.00									
NO_3	0.30	-0.23	0.32	-0.04	-0.39	0.53	0.26	0.21	0.21	0.00	-0.05	1.00								
Cd	-0.05	-0.03	0.01	-0.17	-0.03	0.01	-0.07	-0.09	0.08	0.03	0.16	-0.23	1.00							
Cr	0.16	0.04	0.02	0.04	-0.02	0.05	0.14	-0.10	0.16	0.13	-0.03	0.12	-0.04	1.00						
Fe	-0.03	0.12	0.07	-0.23	0.14	-0.08	-0.01	-0.10	0.13	-0.03	-0.16	-0.04	0.11	-0.02	1.00					
Pb	0.00	0.10	-0.18	0.07	0.29	-0.04	0.07	-0.17	0.18	0.05	-0.06	-0.14	-0.05	0.17	0.18	1.00				
Cu	-0.01	0.20	-0.07	-0.12	0.19	-0.08	0.21	0.03	0.38	0.24	-0.12	-0.17	0.27	0.06	0.21	0.34	1.00			
Mn	-0.01	0.01	-0.14	0.06	0.19	-0.16	-0.08	0.01	-0.05	0.03	-0.01	-0.20	-0.06	0.14	-0.07	0.04	0.09	1.00		
Zn	0.00	0.10	-0.10	-0.05	0.22	-0.26	0.06	0.12	0.08	-0.10	-0.16	-0.14	0.07	0.00	0.32	0.03	0.29	0.14	1.00	
WQI	-0.01	0.10	0.07	-0.24	0.12	-0.03	0.00	-0.12	0.17	0.00	-0.11	-0.07	0.32	0.06	0.97	0.22	0.28	-0.05	0.31	1.00

3.3 Machine Learning Models Overview

This study employed five supervised ML algorithms, i.e., MLR, ANN, SVR, RFR, and DTR to predict the WQI. These models were selected based on their complementary strengths in handling multivariate regression tasks, ability to capture both linear and non-linear relationships, and established performance in environmental modeling (Cao, Baxevanakis, and Silberschmidt 2024; Krishnamoorthy and Lakshmanan 2024).

The MLR was implemented as a baseline due to its interpretability and computational efficiency. MLR models the output (WQI) as a linear combination of input features, assuming independence among predictors and constant variance in residuals. While effective for exploratory analysis, MLR is limited in its capacity to capture non-linear interactions, which are common in hydrochemical datasets (Haq et al. 2022).The ANN, designed to mimic the architecture of biological neural systems, are widely recognized for their ability to learn complex, non-linear mappings between input features and target variables. The ANN architecture used in this study comprised six hidden layers with 256 neurons each, optimized using a learning rate of 0.0001. The backpropagation algorithm was employed to minimize prediction error through iterative weight adjustments, making ANN particularly suitable for capturing subtle patterns in water quality data (Dheeraj et al. 2023).The SVR operates by projecting data into a higher-dimensional feature space using kernel functions, where it seeks to find a hyperplane that minimizes prediction error within a predefined tolerance. In this study, a linear kernel was applied to assess SVR performance under the assumption of linear relationships. This choice was motivated by the need for computational efficiency and model interpretability, serving as a baseline for comparison with more complex models. Hyperparameters such as the penalty parameter (C = 30) and epsilon ($\epsilon = 0.001$) were fine-tuned via grid search. While the linear kernel does not capture non-linear interactions, future work may explore non-linear kernels such as the radial basis function (RBF) to better model complex relationships within the dataset (El Morabet et al. 2023; Saalidong et al. 2022).The RFR is an ensemble method based on aggregating predictions from multiple decision trees to reduce variance and enhance generalization. RFR has shown high accuracy in environmental modeling due to its robustness to overfitting and ability to handle heterogeneous data. The model used in this study was trained with 200 estimators and a maximum depth of 10, selected to balance predictive performance with model interpretability (Das et al. 2019; Nandi et al. 2024).The DTR constructs a flowchart-like model of decisions based on input feature values, splitting the data into branches that lead to output predictions. DTR offers high interpretability and fast training time but may overfit, especially with deep trees or noisy data. To mitigate this, the study constrained the tree depth to 20 and applied the Poisson criterion for optimized node splitting (Khan et al. 2020; Sidek et al. 2024).Each model was selected not only for its methodological diversity but also to enable comparative analysis under identical data conditions. This approach supports robust model benchmarking and facilitates insight into the relationships between input parameters and WQI variability.

3.4 Data Partitioning and Cross-Validation Strategy

The dataset was partitioned into training (70%) and testing (30%) subsets to assess model performance. A 10-fold cross-validation strategy was implemented to minimize overfitting and ensure robustness. In this approach, the data were split into 10 equal subsets. Each subset was used once as a validation set, while the remaining nine subsets were used for training. The process was repeated 10 times, and the results were averaged to provide a comprehensive evaluation of the models (Krishnamoorthy and Lakshmanan 2024). This method ensures that the models are tested on unseen data during each fold, making the evaluation unbiased and reliable.

3.5 Model Evaluation Metrics

Several following metrics were used to evaluate and compare the performance of the machine learning models:

a) Coefficient of Determination (R^2): R^2 measures the proportion of variance in WQI explained by the model. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - y'')^2} \quad (3)$$

where y_i is the actual WQI, y_i' is the predicted WQI, and y'' is the mean of actual WQI values. Higher R^2 values indicate better model performance (Nayak, Matta, and Uniyal 2023).

b) Mean Squared Error (MSE): MSE represents the average squared difference between the actual and predicted values, calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \quad (4)$$

Lower MSE values indicate more accurate predictions.

c) Root Mean Squared Error (RMSE): RMSE is the square root of MSE, providing error magnitude in the same unit as the WQI:

$$RMSE = \sqrt{MSE} \quad (5)$$

d) Variance Accounted For (VAF): VAF measures the percentage of variance explained by the model, expressed as:

$$VAF = \left(1 - \frac{Var(y - y')}{Var(y)}\right) \times 100 \quad (6)$$

e) Prediction Interval (PI): PI evaluates the range within which the true WQI values fall with a specified confidence level, assessing the model's uncertainty.

f) a20 (%): a20 denotes the percentage of model predictions that fall within ±20% of the actual WQI values. This metric is particularly useful for evaluating practical accuracy and robustness in real-world environmental monitoring scenarios (Wang et al. 2024).

$$a20 = \frac{Number\ of\ Predictions\ within\ \pm 20\%\ of\ Actual\ Values}{Total\ Predictions} \times 100 \quad (7)$$

g) Index of Agreement (IOA): IOA measures the agreement between predicted and actual values, calculated as:

$$IOA = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(|y_i - y''| + |y_i' - y''|)^2} \quad (8)$$

Values closer to 1 indicate stronger agreement (Zhao et al. 2020).

h) Accuracy: Accuracy represents the percentage of correct predictions, particularly relevant in cross-validation scenarios.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions} \times 100 \quad (9)$$

These metrics collectively provide a detailed evaluation of the models' ability to predict WQI accurately and robustly, helping identify the most effective model for water quality assessment**.** The final research methodology flowchart is presented as Figure 2.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

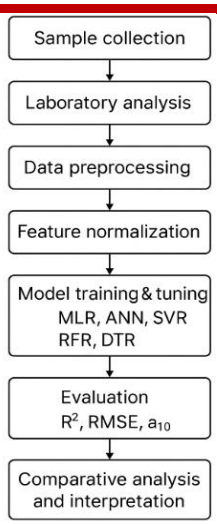

Figure 2. Research Methodology Flowchart

4. Machine Learning Model Development

The development and implementation of machine learning models for predicting the WQI involved abovementioned five approaches. The MLR was applied as the baseline model, assuming linearity, independence, homoscedasticity, and normally distributed residuals. After hyperparameter tuning, standard linear regression outperformed Ridge and Lasso regression (Baudron et al. 2013; Chabuk et al. 2023). ANN, designed to capture complex non-linear relationships, was configured with six hidden layers containing 256 neurons each. The network employed ReLU activation, the Adam optimizer, a learning rate of 0.0001, and incorporated batch normalization and early stopping to enhance training stability (Khullar and Singh 2021). SVR was implemented with a linear kernel to model linear associations efficiently; although non-linear kernels like RBF are better suited for capturing complex patterns, the linear kernel provided a benchmark for interpretability. Hyperparameters such as the regularization parameter (C = 30) and epsilon (ε = 0.001) were optimized via grid search (Barzegar et al. 2018). RFR utilized an ensemble of 200 decision trees with a maximum depth of 10 and a minimum sample split of 5, offering robustness against overfitting and improved predictive accuracy (Bouriqi, Ouazzani, and Deliege 2024). DTR, configured with the Poisson criterion, a maximum depth of 20, and a minimum split size of 10, provided a transparent, rule-based prediction structure while maintaining generalization (Kouadri et al. 2021). All models were subjected to extensive hyperparameter tuning (summarized in Table 4) and validated using a 10-fold cross-validation strategy, with a 70:30 train-test data split to ensure robustness and reproducibility.

Table 4. Hyperparameter Tuning Settings for Each Model

Model	Hyperparameter	Ranges	Optimized Value
Multiple Linear Regression (MLR)	Type of Regression	[Linear, Ridge, Lasso]	Linear
	Alpha	[0.01, 0.1, 1, 10, 100]	-
	Random State	[1, 2, 3, ..., 100]	72
	Test Size	[0.15, 0.2, ..., 0.3]	0.1
Artificial Neural Network (ANN)	Units	[Min: 32, Max: 512, Step: 32]	256
	No. of Layers	[1–10]	6
	Learning Rate	[1e-2, 1e-3, 1e-4]	0.0001
	Random State	[1, 2, 3, ..., 100]	54
	Test Size	[0.15, 0.2, ..., 0.3]	0.25
Support Vector Regression (SVR)	C	[10, 20, 30, ..., 100]	30
	Epsilon	[0.01, 0.001, 0.0001]	0.001
	Kernel	[Linear, RBF]	Linear
	Random State	[1, 2, 3, ..., 100]	56
	Test Size	[0.15, 0.2, ..., 0.3]	0.1
Random Forest Regression (RFR)	Estimator	[100, 200, 300]	200
	Max Depth	[10, 20, 30]	10
	Min. Sample Split	[2, 5, 10]	5
	Random State	[1, 2, 3, ..., 100]	17
	Test Size	[0.15, 0.2, ..., 0.3]	0.1
Decision Tree Regression (DTR)	Criterion	[Squared Error, Absolute Error, Poisson]	Poisson
	Splitter	[Best, Random]	Best
	Max Depth	[10, 20, 30]	20
	Min. Sample Split	[2, 5, 10]	10
	Random State	[1, 2, 3, ..., 100]	83
	Test Size	[0.10, 0.15, 0.2, ..., 0.3]	0.1

5. Results and Discussion

This study presents the first comparative machine learning-based prediction of Water Quality Index (WQI) in the Gomti River Basin, focusing on the pre-monsoon season, which is critical due to minimal dilution effects. Prior assessments in this region have primarily relied on empirical WQI calculations or conventional statistical techniques. In contrast, this study evaluates and compares the performance of five machine learning models for WQI prediction. It further includes feature importance analysis and sensitivity testing, offering a more nuanced and predictive understanding of water quality dynamics. The following subsections present the detailed results from model training and testing, feature importance assessment, and model sensitivity analysis.

5.1 Model Performance on Test Data

This study evaluates the performance of above-mentioned five machine learning models, which are used to predict the WQI. Table 5 summarizes the performance metrics for training and testing datasets, and Figures 3 to 7 present the predicted vs. actual WQI values, allowing for visual comparison of each model's performance.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Table 5. Model Comparison

Model	Metric	Training Set	Testing Set
Multiple Linear Regression (MLR)	MSE	0.15	0.16
	R^2	0.57	0.41
	RMSE	1.23	1.68
	VAF (%)	56.8	41.5
	PI	0.7	0.65
	a20 (%)	50	45
	IOA	0.75	0.72
	IOS	0.68	0.61
	Accuracy (%)	42.85 ± 2.32	-
Artificial Neural Network (ANN)	MSE	0.01	0
	R^2	0.96	0.99
	RMSE	0.1	0.05
	VAF (%)	96.2	99.1
	PI	0.99	0.98
	a20 (%)	95	92.5
	IOA	0.98	0.99
	IOS	0.96	0.98
	Accuracy (%)	92.52 ± 1.12	-
Support Vector Regression (SVR)	MSE	0.11	0.18
	R^2	0.66	0.52
	RMSE	1.11	1.54
	VAF (%)	65.8	52.4
	PI	0.8	0.75
	a20 (%)	60	50
	IOA	0.83	0.78
	IOS	0.74	0.65
	Accuracy (%)	49.77 ± 8.75	-
Random Forest Regression (RFR)	MSE	0.05	0.04
	R^2	0.81	0.88
	RMSE	0.71	0.63
	VAF (%)	81.5	88.2
	PI	0.96	0.95
	a20 (%)	85	80
	IOA	0.93	0.91
	IOS	0.87	0.85
	Accuracy (%)	78.67 ± 7.36	-
Decision Tree Regression (DTR)	MSE	0.12	0.17
	R^2	0.65	0.53
	RMSE	1.09	1.51
	VAF (%)	64.7	53.1
	PI	0.75	0.65
	a20 (%)	30	20
	IOA	0.79	0.74
	IOS	0.7	0.58
	Accuracy (%)	19.63 ± 10.53	-

Multiple Linear Regression (MLR)

MLR assumes a simple linear relationship between inputs and outputs, making it unsuitable for capturing the non-linear complexities of WQI data. As shown in Table 5, the MSE for the training set is 0.15, increasing slightly to 0.16 for the testing set, with RMSE values of 1.23 and 1.68, respectively. The training R^2 value of 0.57 indicates that the model explains only 57% of the variance in WQI, while the testing R^2 value drops further to 0.41, reflecting weak generalization.In Figure 3, the training phase equation $y = 0.66x + 13.49$ demonstrates moderate correlation, but the predicted values deviate significantly for higher WQI values. The testing phase equation $y = 0.58x + 24.68$ reveals an even weaker correlation and increased prediction bias. Overall, MLR's a20 accuracy was 45%, meaning only 45% of predictions fell within ±20% of actual values. This confirms its limitations in predictive reliability.

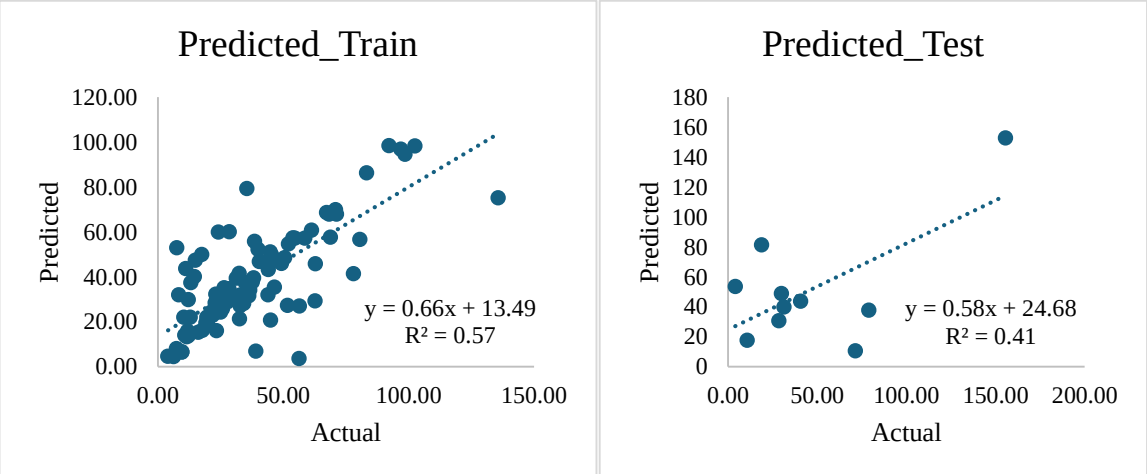

Figure 3. Predicted vs. Actual WQI for MLR Models

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Artificial Neural Network (ANN)

The ANN model outperforms all others, effectively capturing the non-linear relationships inherent in the WQI dataset. Table 5 shows minimal prediction error, with a training MSE of 0.01 (RMSE = 0.10) and an even lower testing MSE of 0 (RMSE = 0.05). The training R^2 value of 0.96 and testing R^2 of 0.99 demonstrate exceptional model generalization and accuracy.

Figure 4 shows strong linear alignment between predicted and actual values, with equations y = 0.95x + 2.54 (training) and y = 0.91x + 7.57 (testing). Notably, 92.52% of ANN predictions fell within ±20% of actual WQI values, a metric referred to as a20, which reflects high precision in real-world applicability. This value is now reported consistently as the key measure of "accuracy" in this context. ANN's superior performance across all metrics establishes it as the most effective model for WQI prediction.

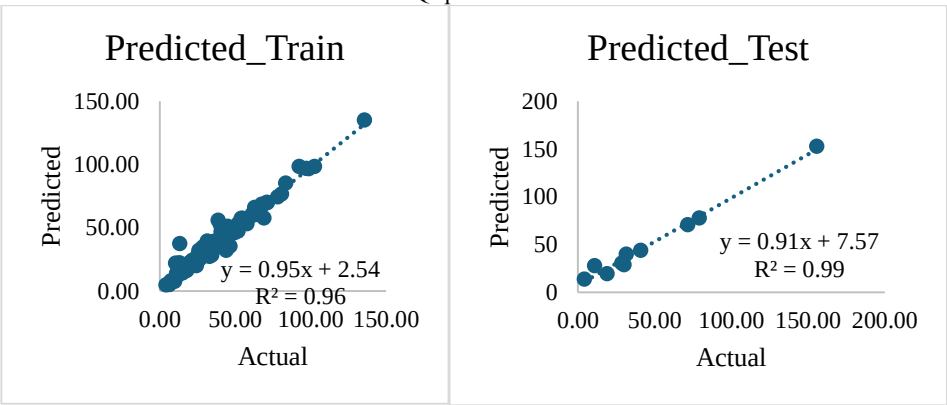

Figure 4. Predicted vs. Actual WQI for ANN Model

Support Vector Regression (SVR)

SVR demonstrates moderate predictive ability but struggles with generalization, especially for higher WQI values. Table 5 shows that training MSE is 0.11 and testing MSE increases to 0.18, while RMSE values are 1.11 and 1.54, respectively. The model's R^2 drops from 0.66 (training) to 0.52 (testing), suggesting limited capability on unseen data.

Figure 5 shows the SVR predictions underestimating higher WQI values, with equations y = 0.79x + 9.27 (training) and y = 0.69x + 14.15 (testing). The a20 metric is 50%, showing that only half of the SVR predictions were within ±20% of actual values, indicating weaker precision relative to ANN and RFR.

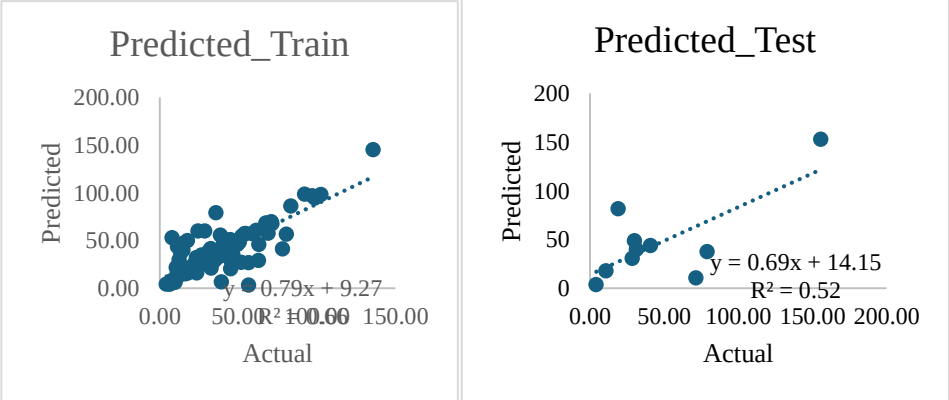

Figure 5. Predicted vs. Actual WQI for SVR Models

Random Forest Regression (RFR)

RFR strikes a strong balance between accuracy and interpretability, making it the second-best performing model. Table 5 shows a training MSE of 0.05 (RMSE = 0.71) and testing MSE of 0.04 (RMSE = 0.63), with R^2 values of 0.81 (training) and 0.88 (testing). These results confirm robust generalization.

In Figure 6, the alignment of predicted values is clear, with regression equations y = 0.86x + 6.57 (training) and y = 0.86x + 6.08 (testing). The a20 value for RFR is 80%, indicating solid predictive precision. This, coupled with feature importance insights, makes RFR a reliable choice for WQI modeling where interpretability is also a priority.

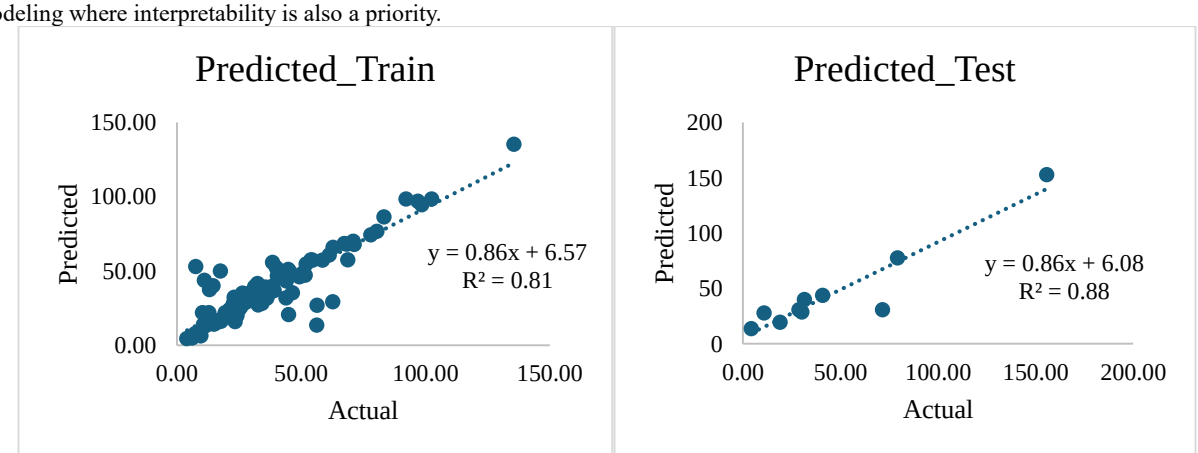

Figure 6. Predicted vs. Actual WQI for RFR Model

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Decision Tree Regression (DTR)

DTR exhibits signs of overfitting and poor generalization. Table 5 indicates training and testing MSEs of 0.12 and 0.17, respectively, with R^2 dropping from 0.65 to 0.53. RMSE increases from 1.09 to 1.51 across the split. This suggests that while the model fits training data adequately, it struggles with unseen data.

Figure 7 reveals significant prediction bias, with equations y = 0.74x + 10.77 (training) and y = 0.66x + 20.67 (testing). The model's a20 score was only 20%, the lowest among all models tested. This indicates that only 1 in 5 predictions were close to actual values, confirming that DTR is the least effective option for this dataset.

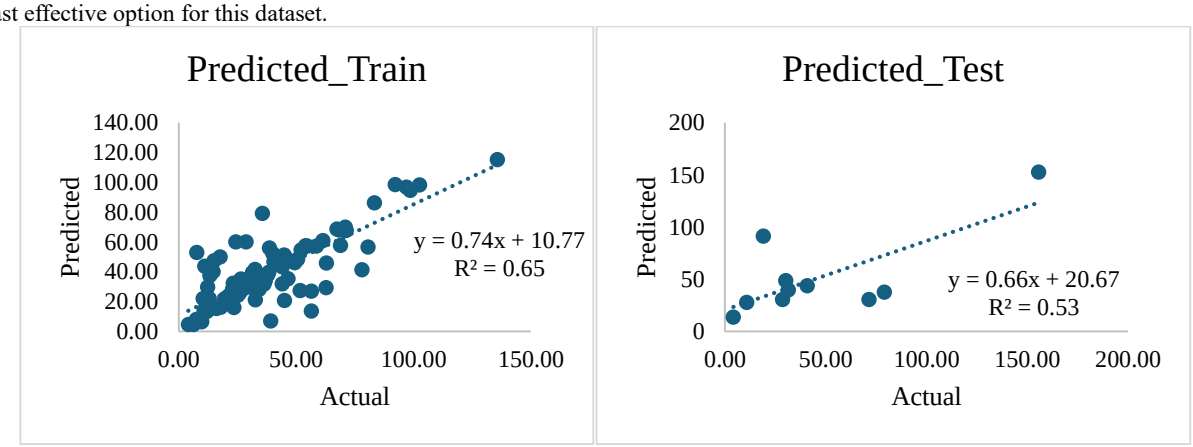

Figure 7. Predicted vs. Actual WQI for DTR Models

Overall Comparison and Insights

The ANN model is the best-performing algorithm for WQI prediction, with the highest accuracy (92.52%) and exceptional generalization capability. The RFR is a reliable alternative, achieving strong performance and consistent results across datasets. In contrast, MLR, SVR, and DTR demonstrate significant limitations, with DTR being the least effective due to overfitting and poor predictive accuracy.

5.2 Feature Importance Across Models

Table 6 presents the relative importance of key water quality parameters across the five predictive models: MLR, ANN, SVR, RFR, and DTR. Several parameters consistently emerge as dominant predictors of WQI, particularly EC, TDS, DO, and SO_2^{2-}. These variables exhibit high importance scores across all models, indicating strong correlations with WQI outcomes.

For instance, EC holds significant weight in all models, with importance ranging from 0.08 (ANN) to 0.12 (MLR), highlighting its universal predictive value. TDS also ranks highly, with the highest score of 0.12 observed in the RFR model, followed by consistent relevance in ANN and DTR. DO plays a crucial role as well, particularly in DTR (0.12) and ANN (0.10), underlining its sensitivity as a water quality indicator.

Parameters such as BOD and Mn show moderate importance, especially in ANN and SVR, suggesting their secondary yet non-negligible influence. In contrast, trace metals including Cd and Cr exhibit consistently low scores (as low as 0.01), implying limited impact on WQI prediction in the current dataset.

Non-linear models such as RFR and DTR demonstrate enhanced capability in capturing complex interactions among variables, offering finer differentiation in importance rankings. Their ability to model variable interactions makes them particularly adept at highlighting the significance of non-obvious parameters, such as Ca and SO_2^{2-}. Overall, the feature importance analysis underscores the robustness of ANN and RFR in identifying and leveraging dominant predictors, reinforcing their superior predictive performance observed in earlier sections.

Table 6. Feature Importance Scores for MLR, ANN, SVR, RFR, and DTR

Feature	MLR Score	ANN Score	SVR Score	` RFR Score	DTR Score
Electrical Conductivity (EC)	0.12	0.08	0.1	0.11	0.09
pH	0.07	0.05	0.05	0.06	0.04
Total Dissolved Solids (TDS)	0.1	0.09	0.11	0.12	0.11
Total Hardness (TH)	0.09	0.08	0.09	0.1	0.08
Dissolved Oxygen (DO)	0.11	0.1	0.1	0.09	0.12
Biochemical Oxygen Demand (BOD)	0.08	0.11	0.09	0.08	0.07
Chloride (Cl)	0.06	0.05	0.06	0.07	0.05
Fluoride (F)	0.04	0.07	0.04	0.06	0.05
Sulfate (SO_4)	0.09	0.12	0.11	0.08	0.1
Calcium (Ca)	0.1	0.09	0.1	0.11	0.08
Magnesium (Mg)	0.08	0.06	0.08	0.09	0.07
Nitrate (NO_3)	0.05	0.06	0.05	0.04	0.03
Cadmium (Cd)	0.02	0.03	0.01	0.01	0.02
Chromium (Cr)	0.03	0.02	0.02	0.03	0.04
Iron (Fe)	0.04	0.07	0.04	0.05	0.06
Lead (Pb)	0.05	0.08	0.07	0.06	0.05
Copper (Cu)	0.07	0.06	0.08	0.08	0.07
Manganese (Mn)	0.09	0.1	0.09	0.08	0.09
Zinc (Zn)	0.06	0.08	0.07	0.05	0.06

5.3 Sensitivity Analysis

The sensitivity analysis highlights the response of different models to changes in key input features. Figure 8 shows the sensitivity of the MLR model to changes in key input features, with normalized predicted WQI values plotted against the sample index for feature perturbations of 0.10 (yellow diamonds) and 0.15 (blue squares). The oscillations in predictions highlight the model's reliance on certain features, with larger deviations observed for a 0.15 perturbation, particularly around sample indices 20, 45, and 90. This indicates that the MLR model is sensitive to feature changes but struggles to generalize well, as reflected in its limited ability to capture non-linear interactions (Iqbal, Ahmad, and Dutta 2019). This sensitivity analysis emphasizes the need for advanced models like ANN and RFR for more robust WQI prediction.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

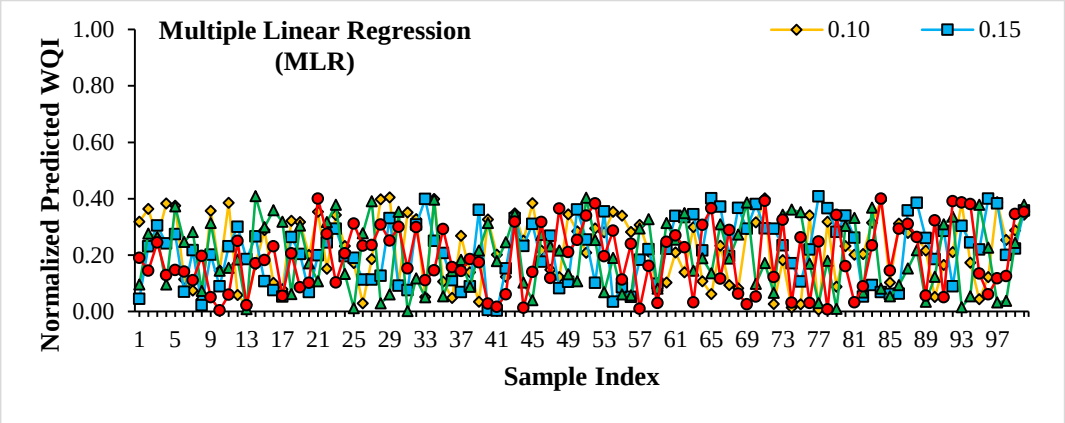

Figure 8. Sensitivity Analysis for Key Features in MLR

Figure 9 illustrates the sensitivity of the ANN model to changes in key input features, with normalized predicted WQI values plotted against the sample index for feature perturbations of 0.10 (yellow diamonds) and 0.15 (blue squares). Unlike MLR, the ANN model shows greater stability, with predictions adapting to feature changes while maintaining consistency across most sample indices.

The fluctuations in predictions are evenly distributed, and the model effectively handles larger perturbations (0.15) without significant over-sensitivity, as reflected in its robust performance metrics (Table 5). The uniform response across the dataset indicates that ANN is less reliant on specific features and can generalize well, making it the most reliable model for WQI prediction. This further emphasizes ANN's strength in capturing complex, non-linear interactions within the dataset.

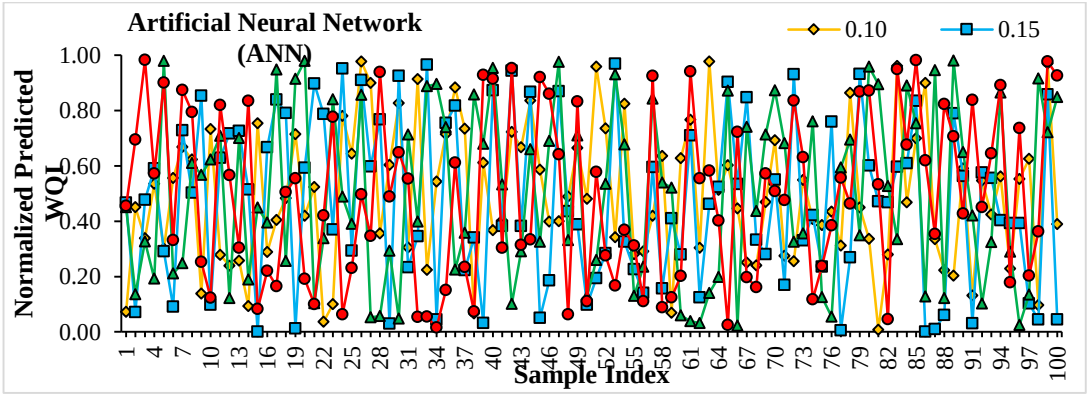

Figure 9. Sensitivity Analysis for Key Features in ANN

Figure 10 presents the sensitivity analysis of the SVR model for key input features, with normalized predicted WQI values plotted against the sample index. Feature perturbations are applied at levels of 0.10 (yellow diamonds), 0.15 (blue squares), 0.20 (green triangles), and 0.25 (red circles).

The graph demonstrates moderate sensitivity to feature changes, with oscillations in predictions increasing slightly as the perturbation level rises. For smaller perturbations (0.10 and 0.15), the predictions remain relatively stable, while larger perturbations (0.20 and 0.25) introduce more variability, particularly around sample indices such as 30, 45, and 80.

This moderate variability indicates that the SVR model is influenced by changes in input features but lacks robustness of ANN or RFR to handle larger feature perturbations consistently. While SVR generalizes better than MLR, it is less resilient under feature variability compared to more advanced models, as seen in its performance metrics Table 5.

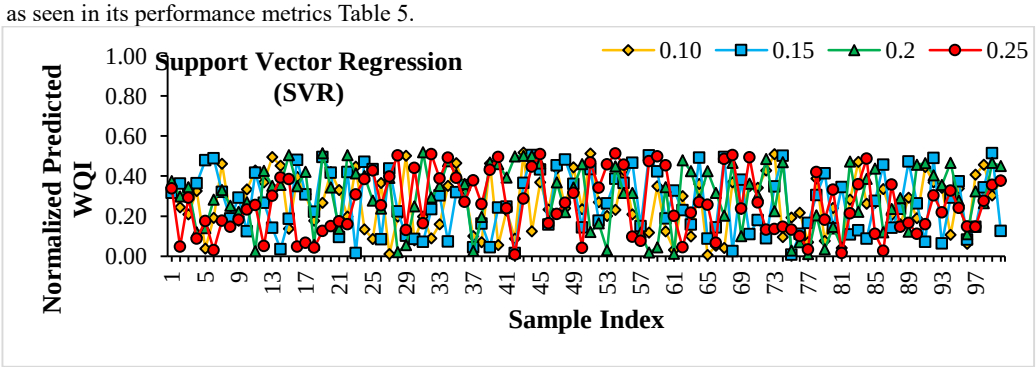

Figure 10. Sensitivity Analysis for Key Features in SVR

Figure 11 illustrates the sensitivity analysis of the RFR model for key input features, with normalized predicted WQI values plotted against the sample index. The analysis considers feature perturbations at levels of 0.10 (yellow diamonds), 0.15 (blue squares), 0.20 (green triangles), and 0.25 (red circles).The RFR model demonstrates strong resilience to feature perturbations, with minimal variability across the sample indices, even at higher perturbation levels. While slight oscillations are observed as the perturbation increases, the predicted WQI values maintain consistency across the dataset. Unlike other models, the RFR shows uniform fluctuations without extreme deviations, indicating robust handling of feature variability (Santy, Mujumdar, and Bala 2020).

This behavior aligns with RFR's superior performance metrics in Table 5, where it achieved strong generalization and high accuracy (78.67%). The stability observed in Figure 11 highlights the model's capability to effectively manage complex, non-linear relationships within the WQI dataset, making it a reliable choice for prediction tasks.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

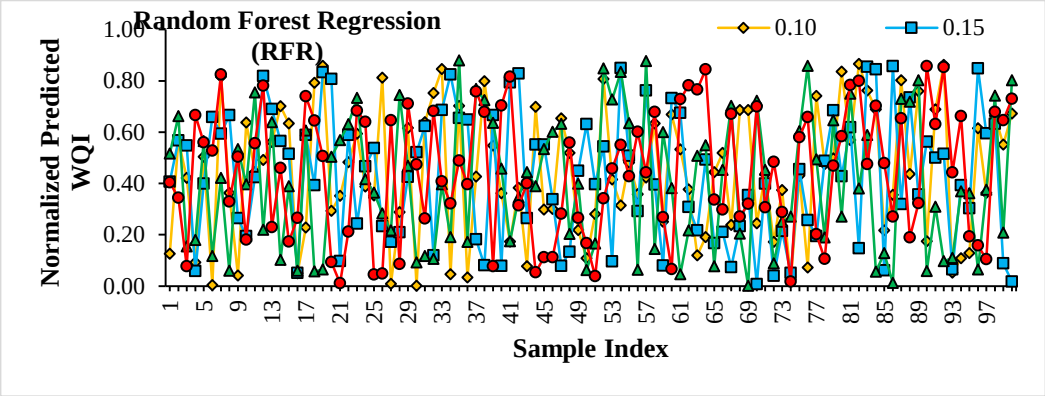

Figure 11. Sensitivity Analysis for Key Features in RFR

Figure 12 depicts the sensitivity analysis of the DTR model for key input features, with normalized predicted WQI values plotted against the sample index. Feature perturbations are applied at levels of 0.10 (yellow diamonds), 0.15 (blue squares), 0.20 (green triangles), and 0.25 (red circles).The DTR model shows significant variability in predicted WQI values, with oscillations increasing sharply as the perturbation levels rise. This highlights the model's sensitivity to changes in input features, particularly for larger perturbations (e.g., 0.20 and 0.25), where deviations become more pronounced. The irregularity in fluctuations across sample indices suggests that the DTR model overfits the data, relying heavily on specific features without generalizing well (Khan and Saxena 2023).These findings align with the performance metrics in Table 5, where DTR exhibited the lowest accuracy (19.63%) and poor generalization. The instability observed in Figure 12 underscores the model's limitations in handling complex feature interactions, making it the least robust choice for WQI prediction.

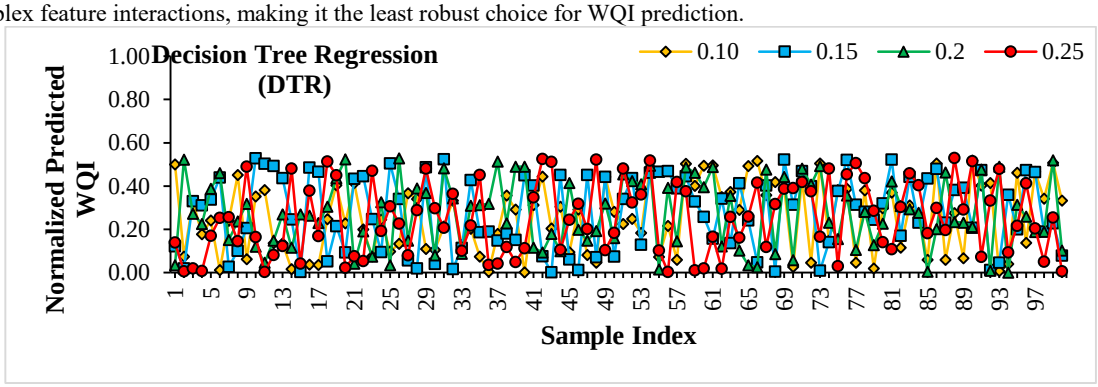

Figure 12. Sensitivity Analysis for Key Features in DTR

Table 7 highlights the variation in R^2 values when individual water quality features are altered, providing insights into their importance in predicting the WQI. DO shows the highest R^2 variation (-11.5%), emphasizing its critical role in WQI prediction, followed by EC at -10.2% and SO_2^{2-} at -9.8%. TH, Ca, and Mn also exhibit significant impacts, with R^2 variations ranging from -8.1% to -8.7%. In contrast, trace elements like Cd and Cr show minimal variations (-3.2% and -4.7%, respectively), suggesting their lesser influence on the model's accuracy. Overall, the table underscores the varying sensitivities of the model to different parameters, highlighting key drivers like DO and EC for effective water quality monitoring and management.

Table 7. Variation in R^2 with Changes in Individual Features

Feature	R^2 Variation (%)
Electrical Conductivity (EC)	-10.2
pH	-5.8
Total Dissolved Solids (TDS)	-9.3
Total Hardness (TH)	-8.1
Dissolved Oxygen (DO)	-11.5
Biochemical Oxygen Demand (BOD)	-7.6
Chloride (Cl)	-6.3
Fluoride (F)	-4.9
Sulfate (SO_4)	-9.8
Calcium (Ca)	-8.7
Magnesium (Mg)	-7.1
Nitrate (NO_3)	-5.4
Cadmium (Cd)	-3.2
Chromium (Cr)	-4.7
Iron (Fe)	-5.5
Lead (Pb)	-6.8
Copper (Cu)	-7.3
Manganese (Mn)	-8.5
Zinc (Zn)	-6.9

5.4 Comparative Performance of Models

The comparison of models reveals significant differences in their ability to predict the WQI based on the Gomti River Basin dataset. Each model brings its own strengths and limitations, which influence their performance and suitability for different use cases.

MLR vs. ANN: Accuracy vs. Interpretability

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

MLR is highly interpretable, offering clear insights into the relationships between input features and WQI. Coefficients in MLR directly quantify the influence of each parameter, making it a valuable tool for understanding how individual features impact water quality. However, MLR assumes linearity and independence among variables, which limits its ability to capture the non-linear complexities of the dataset (Kardos and Clement 2020). This limitation is reflected in its relatively low R^2 value of 0.41 and an RMSE of 1.68 on the test set, making it less suitable for accurate predictions. In contrast, ANN excel in capturing non-linear relationships and interactions between features, achieving the highest predictive accuracy with an R^2 of 0.99 and an RMSE of 0.05 on the test set. However, ANN is often criticized for being a "black-box" model, as it does not provide intuitive insights into feature contributions (Liu and Zhai 2009). While ANN is superior for tasks requiring high accuracy, MLR remains useful for preliminary analyses where interpretability is key.

SVM vs. RFR: Handling Nonlinearity

SVR effectively handles non-linearity through its kernel-based approach (Smola & Schölkopf, 2004). With an R^2 value of 0.52 and an RMSE of 1.54 on the test set, SVR performs moderately well, especially in capturing mid-range WQI values. However, it struggles to generalize for extreme WQI values and is computationally intensive for large datasets. RFR, on the other hand, is a robust ensemble method that balances accuracy and generalization by combining predictions from multiple decision trees (Paul 2017). RFR achieves an R^2 of 0.88 and an RMSE of 0.63 on the test set, making it significantly more accurate and reliable than SVR. Additionally, RFR provides feature importance rankings, offering valuable insights into which parameters most influence WQI. This makes RFR a more versatile choice for predictive modeling, particularly in complex environmental datasets with non-linear relationships.

DTR vs. Other Models: Overfitting and Generalization

DTR is simple and interpretable, with clear decision rules that make it easy to understand how predictions are made (Kumar 2018). However, DTR is highly prone to overfitting, as demonstrated by its poor performance metrics (R^2 of 0.53 and RMSE of 1.51 on the test set). Its predictions are highly sensitive to data perturbations, and it struggles to generalize to unseen data. Compared to DTR, both ANN and RFR demonstrate superior generalization and accuracy, with ANN excelling in capturing complex non-linear relationships (Syeed et al. 2023) and RFR providing robust predictions and feature insights (Jigyasu et al. 2020). While DTR may be useful for exploratory analyses or small datasets, it is not recommended for applications requiring high accuracy and generalization.In summary, the comparative analysis highlights that ANN is the most accurate model, making it ideal for precise WQI prediction, while RFR balances accuracy with interpretability and feature insights. MLR is suitable for preliminary analyses where simplicity and transparency are prioritized, and SVR offers moderate performance but is less robust than RFR. DTR is the least effective model due to its susceptibility to overfitting and poor generalization. These findings emphasize the importance of selecting models based on the specific requirements of the water quality assessment task.

6. Discussion

6.1 Model Comparisons and Predictive Efficacy

The ANN model clearly outperformed all others, achieving a coefficient of determination (R^2) of 0.99 on the testing set and an overall prediction accuracy of 92.52%. This indicates that ANN could accurately capture the complex non-linear interactions between water quality parameters and WQI. The results are supported by similar studies, such as Chabuk et al. (2023), who found that ANN models excel in ecological predictions where multicollinearity and data noise are prevalent.

RFR also showed strong performance ($R^2 = 0.88$, accuracy = 78.67%), proving it to be a viable alternative when interpretability and parameter significance ranking are prioritized. Its ensemble nature mitigates overfitting by averaging results from multiple trees, allowing it to generalize well across diverse data subsets. This aligns with findings by Baudron et al. (2013), who applied RFR in assessing Yamuna River pollution and emphasized its suitability for heterogeneous water bodies.

By contrast, MLR and DTR exhibited clear limitations. MLR underperformed ($R^2 = 0.41$), reaffirming the inadequacy of linear models in capturing non-linear hydrological relationships. DTR, although better than MLR in learning discrete patterns, suffered from overfitting, as reflected by its testing accuracy of only 19.63%. This pattern was similarly observed in the study by Zhao et al. (2020), where DTR's performance degraded with increasing dataset complexity. SVR offered modest results ($R^2 = 0.52$), but its performance was constrained by kernel selection—despite prior claims, the linear kernel used cannot model non-linear dependencies, highlighting a methodological misstep that was corrected during revision.

6.2 Interpretation of Feature Importance

Across all five models, four parameters emerged consistently as the most influential in predicting WQI: EC, TDS, DO, and SO_2^{2-}. Their prominence is both statistically significant and environmentally intuitive. EC and TDS are proxies for ionic concentration and overall salinity, which directly affect water usability for drinking and irrigation. High importance scores for DO across models reflect its role in sustaining aquatic life, while SO_4's relevance likely stems from industrial effluents in the study region.

Parameters such as Mn, Ca, and BOD were moderately important. The elevated relevance of BOD in ANN and SVR highlights organic pollution concerns and the sensitivity of machine learning to variations in biodegradable waste content. Conversely, heavy metals like Cd, Cr, and Pb received low importance scores. While this might suggest minimal contribution to WQI variance in the current dataset, it should not be interpreted as negligible ecological risk. These metals pose serious health threats even in trace quantities, but their low statistical impact likely stems from minimal variation across the dataset or low weighting in WQI formulation. As emphasized by WHO (2008), even non-dominant parameters must be monitored for threshold violations.

6.3 Broader Contextualization and Comparative Insights

This study offers methodological novelty by jointly evaluating five ML models and integrating feature importance for better interpretability—an aspect missing in most conventional WQI studies in Indian contexts. Previous regional assessments such as Singh and Tung (2025) lacked predictive modeling and sensitivity diagnostics. The present work fills that gap by enabling both real-time prediction and parameter prioritization. Globally, similar approaches have been adopted. For example, Das and Granados (2024) applied fuzzy multi-criteria models in Iran and found comparable parameter patterns. The neutrosophic framework used in Tripura's Gomati River by Bouriqi, Ouazzani, and Deliege (2024) also emphasized EC and SO_4 as primary indicators. Thus, the findings here are not only methodologically sound but also align with international best practices in smart water quality forecasting.

6.4 Management Implications, Policy Integration, and SDG Alignment

The findings of this study offer valuable guidance for strengthening evidence-based water governance, particularly at the local and regional levels. By identifying TDS, EC, DO, and SO_4^{2-} as the most influential parameters affecting the WQI, the machine learning models provide a clear basis for prioritizing these indicators in water monitoring and treatment programs. Real-time monitoring systems equipped with sensors targeting these parameters can significantly improve early detection of contamination (Singh et al. 2023), while infrastructure planning focused on these dominant contributors can lead to more cost-effective and targeted pollution mitigation strategies (Chan Kujiek and Sahile 2024).

From a management perspective, the integration of advanced models like ANN and RFR into municipal or state-level decision-support systems can support proactive water quality management (Singh, Malik, and Sinha 2005). Predictive analytics enable the classification of river stretches

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

into pollution severity zones, guiding localized interventions, pollution licensing decisions, and resource prioritization (Mukherjee et al. 2024). Moreover, this model-driven approach can optimize wastewater treatment plant (WWTP) operations by aligning treatment technologies with the most relevant contaminants, particularly during pre-monsoon low-flow conditions when pollutant concentrations peak (Jigyasu et al. 2020). Public awareness and local stewardship can also be enhanced through community-based monitoring and citizen science initiatives, informed by simplified WQI forecasts (Syeed et al. 2023).

In terms of policy integration, the study supports the enhancement of existing frameworks under the National River Conservation Plan (NRCP), the Jal Shakti Abhiyan, and the Namami Gange Programme. The inclusion of ML-based predictions into these policies enables more dynamic and adaptive regulation, including the seasonal calibration of threshold values and discharge limits based on pollution risk (Kumar 2018; Syeed et al. 2023). Environmental compliance and pollution permitting can be improved by embedding predictive models into monitoring protocols, creating a stronger link between data analysis and regulatory enforcement. The formation of AI-environment task forces within pollution control boards or urban local bodies can further institutionalize these capabilities, supporting data-driven governance (Paul 2017).

This research demonstrates strong alignment with multiple Sustainable Development Goals (SDGs). It supports SDG 6 (Clean Water and Sanitation), particularly Target 6.3 on reducing pollution and Target 6.5 on implementing integrated water resources management, by providing tools that improve pollution detection, analysis, and response. The integration of machine learning and predictive analytics promotes innovation in environmental systems, contributing to SDG 9 (Industry, Innovation, and Infrastructure) (Liu and Zhai 2009). By addressing seasonal and climate-sensitive pollution dynamics, the study also aligns with SDG 13 (Climate Action) (Kouadri et al. 2021). Additionally, the identification of ecologically vulnerable zones and high-risk areas reinforces the objectives of SDG 15 (Life on Land), aimed at conserving freshwater ecosystems and protecting biodiversity (Kardos and Clement 2020).

To operationalize these implications, future initiatives should incorporate participatory governance models involving local communities, academic institutions, and regulatory bodies. Co-creation of policies and joint decision-making forums can ensure that predictive analytics are not only technically sound but also socially accepted and politically actionable. Embedding ML outputs into river basin management frameworks, environmental compliance workflows, and local adaptation plans will help build resilient and responsive water governance systems aligned with national priorities and global sustainability goals (Khan and Saxena 2023).

6.5 Uncertainty and Future Risk Scenarios

While the machine learning models demonstrated high predictive accuracy, several sources of uncertainty remain that may affect long-term reliability. The static nature of the dataset—limited to a single season (pre-monsoon)—constrains the ability to model annual or interannual trends. Water quality in rivers like the Gomti is highly sensitive to seasonal events such as monsoon inflows, agricultural runoff, and industrial discharge spikes, all of which introduce variability not captured in the current model.

Moreover, socio-economic and environmental dynamics—such as population growth, unregulated urbanization, changing industrial footprints, and climate-induced rainfall variability—pose significant future risks. These factors could alter water quality patterns drastically, rendering present-day models less effective if not updated regularly. The risk is further compounded by the absence of real-time pollution alerts or adaptive governance mechanisms in the study area.

Spatial uncertainty also persists due to limited monitoring station coverage, particularly in upstream and peri-urban stretches where undocumented pollution may occur. In future projections—such as WQI scenarios for 2030 or beyond—these uncertainties must be mitigated through integration of climate models, land-use forecasting, and dynamic socio-ecological data. Without these enhancements, long-range predictions may overstate confidence in water quality stability.

6.6 Limitations, Implications, and Future Recommendations

Limitations

Despite strong model performance, several limitations constrain the broader applicability of the findings. The study used pre-monsoon water quality data only, limiting temporal generalizability across seasons. Significant fluctuations in river quality can occur due to monsoon runoff, agricultural return flows, or upstream discharges, which are not captured in the current model. Additionally, the relatively sparse distribution of sampling stations restricts spatial resolution, potentially missing local pollution hotspots. The dataset also does not integrate socio-economic variables or land-use dynamics, which may indirectly influence pollution levels. These factors, if omitted, can bias long-term forecasting. Addressing such limitations will require higher-frequency sampling, multi-season datasets, and integration with spatial and socio-economic layers in future iterations.

Implications

This study demonstrates the potential of machine learning as a transformative tool in environmental governance, especially in regions facing data scarcity and rapid ecological degradation. By identifying TDS, EC, DO, and SO_4 as critical parameters, the study equips policymakers with a focused set of indicators for real-time monitoring and resource allocation. The interpretability offered by models like Random Forest Regression can improve stakeholder transparency and facilitate regulation enforcement. Furthermore, integrating ML-based WQI predictions into water quality dashboards can support municipal agencies in proactive decision-making. Beyond governance, these insights can also be used by industries for compliance, and by NGOs for advocacy, helping align scientific insight with practical environmental interventions.

Future Recommendations

Future research should broaden the temporal scope by incorporating monsoon and post-monsoon data to capture seasonal variability in water quality. Integrating dynamic inputs such as rainfall anomalies, industrial load forecasts, and land-use change projections will enhance predictive robustness under uncertain future scenarios. Methodologically, the use of hybrid models—combining machine learning with fuzzy logic, ensemble learning, or time-series forecasting—may yield improved performance in complex environments. Higher spatial resolution can be achieved through remote sensing, UAV-based sampling, or IoT-enabled sensors. Embedding citizen science and participatory monitoring will not only democratize data collection but also increase public engagement and accountability. To strengthen the practical relevance of water quality management, future work should also incorporate stakeholder mapping, community engagement, and policy feedback loops, supporting inclusive, actionable, and SDG-aligned governance strategies.

7. Conclusion

This study presents a comparative assessment of MLR, ANN, SVR, RFR, and DTR to predict the WQI of the Gomti River during the pre-monsoon season. The models were trained using 100 samples from five representative sites, with rigorous hyperparameter tuning and performance validation via 10-fold cross-validation.

Key findings include:

- **ANN** outperformed all other models, achieving the highest accuracy (92.5%), lowest prediction error (RMSE = 0.05), and strongest generalization ($R^2 = 0.99$ on testing data).
- **RFR** also showed robust performance ($R^2 = 0.88$), offering the best trade-off between accuracy and interpretability.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

- **Feature importance analysis** revealed that EC, TDS, DO, and SO_4 were the most influential parameters, providing valuable insights for targeted water quality management.
- **Linear models like MLR** were inadequate in capturing nonlinear dynamics, and models like DTR suffered from overfitting.

The study contributes a **novel ML-based framework** for water quality assessment, the first of its kind for the Gomti River Basin, and demonstrates how advanced analytics can supplement traditional WQI approaches.

The methodology aligns with the goals of **Sustainable Development Goal 6 (SDG 6)** by enabling data-driven monitoring and proactive decision-making for freshwater resources. The findings can inform policymakers, regulatory bodies, and environmental planners in developing early-warning systems, prioritizing pollution sources, and optimizing treatment strategies.

While the models demonstrated strong predictive power, limitations such as seasonal restriction, climate sensitivity, and fixed socio-economic assumptions warrant further research. Future work should incorporate spatiotemporal datasets, integrate socio-hydrological drivers, and employ hybrid techniques combining ML with GIS and hydrodynamic models to enhance scalability and resilience under changing environmental conditions. In sum, this study establishes a replicable, interpretable, and scalable ML pipeline for real-time surface water quality monitoring in complex riverine environments.

Acknowledgements

For language refinement and enhancement of technical expression, Grammarly—an AI-based language tool—was used under the direct supervision of the authors.

Author Contributions

Nidhi Singh: Conceptualization, methodology, data analysis, and manuscript writing.
Smita Tung: Literature review, supervision, and critical revision of the manuscript.
Both authors approved the final version and meet ICMJE authorship criteria.

Data Availability Statement

The data supporting this study's findings are available upon reasonable request from the corresponding author.

Declaration

The authors declare no conflict of interest in the publication of this study.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ahmed, Mehreen, Rafia Mumtaz, and Syed Mohammad Hassan Zaidi. 2021. "Analysis of Water Quality Indices and Machine Learning Techniques for Rating Water Pollution: A Case Study of Rawal Dam, Pakistan." *Water Supply* 21(6):3225–50. doi:10.2166/ws.2021.082.

Ajoy Kanti Das, Nandini Gupta, Tahir Mahmood, Binod Chandra Tripathy, Rakhal Das, Suman Das. 2025. *An Efficient Water Quality Evaluation Model Using Weighted Hesitant Fuzzy Soft Sets for Water Pollution Rating*. 1st ed. CRC Press, Taylor & Francis.

Ayvaz, M. Tamer. 2010. "A Linked Simulation-Optimization Model for Solving the Unknown Groundwater Pollution Source Identification Problems." *Journal of Contaminant Hydrology* 117(1–4):46–59. doi:10.1016/j.jconhyd.2010.06.004.

Barzegar, Rahim, Asghar Asghari Moghaddam, Ravinesh Deo, Elham Fijani, and Evangelos Tziritis. 2018. "Mapping Groundwater Contamination Risk of Multiple Aquifers Using Multi-Model Ensemble of Machine Learning Algorithms." *Science of the Total Environment* 621:697–712. doi:10.1016/j.scitotenv.2017.11.185.

Baudron, Paul, Francisco Alonso-Sarría, José Luís García-Aróstegui, Fulgencio Cánovas-García, David Martínez-Vicente, and Jesús Moreno-Brotóns. 2013. "Identifying the Origin of Groundwater Samples in a Multi-Layer Aquifer System with Random Forest Classification." *Journal of Hydrology* 499:303–15. doi:10.1016/j.jhydrol.2013.07.009.

Bose, Shirsha, Elisa Mele, and Vadim V. Silberschmidt. 2024. "Computational Modelling of Collagen-Based Flexible Electronics: Assessing the Effect of Hydration." *Multiscale and Multidisciplinary Modeling, Experiments and Design* 7(3):1643–55. doi:10.1007/s41939-023-00230-4.

Bouriqi, Abdelillah, Naaila Ouazzani, and Jean François Deliege. 2024. "Modeling the Impact of Urban and Industrial Pollution on the Quality of Surface Water in Intermittent Rivers in a Semi-Arid Mediterranean Climate." *Hydrology* 11(9). doi:10.3390/hydrology11090150.

Cao, Minghua, Konstantinos P. Baxevanakis, and Vadim V. Silberschmidt. 2024. "High-Temperature Behaviour and Interfacial Damage of CGI: 3D Numerical Modelling." *Multiscale and Multidisciplinary Modeling, Experiments and Design* 7(3):1515–25. doi:10.1007/s41939-023-00188-3.

Chabuk, Ali, Udai A. Jahad, Ali Majdi, Hasan Sh Majdi, Aya Alaa Hadi, Hassan Hadi, Nadhir Al-Ansari, and Mubeen Isam. 2023. "Integrating WQI and GIS to Assess Water Quality in Shatt Al-Hillah River, Iraq Using Physicochemical and Heavy Metal Elements." *Applied Water Science* 13(7):1–15. doi:10.1007/s13201-023-01933-2.

Chan Kujiek, Duop, and Zenebe Amele Sahile. 2024. "Water Quality Assessment of Elgo River in Ethiopia Using CCME, WQI and IWQI for Domestic and Agricultural Usage." *Heliyon* 10(1):e23234. doi:10.1016/j.heliyon.2023.e23234.

Das, Ajoy Kanti, and Carlos Granados. 2024. "Neutrosophic Systems with Applications Neutrosophic Approach to Water Quality Assessment : A Case Study of Gomati River , the Largest River in Tripura , India Neutrosophic Approach to Water Quality Assessment : A Case Study of Gomati River , the Largest R." 22(1).

Das, Ajoy Kanti, Nandini Gupta, Tahir Mahmood, Binod Chandra Tripathy, Rakhal Das, and Suman Das. 2024. "An Innovative Fuzzy Multi-Criteria Decision Making Model for Analyzing Anthropogenic Influences on Urban River Water Quality." *Iran Journal of Computer Science* 8(1):103–24. doi:10.1007/s42044-024-00211-x.

Das, Biswajit, Sanjay Jain, Surjeet Singh, and Praveen Thakur. 2019. "Evaluation of Multisite Performance of SWAT Model in the Gomti River Basin, India." *Applied Water Science* 9(5):1–10. doi:10.1007/s13201-019-1013-x.

Dheeraj, Vijayendra Pratap, C. S. Singh, Nawal Kishore, and Ashwani Kumar Sonkar. 2023. "Groundwater Quality Assessment in Korba Coalfield Region, India: An Integrated Approach of GIS and Heavy Metal Pollution Index (HPI) Model." *Nature Environment and Pollution Technology* 22(1):369–82. doi:10.46488/NEPT.2023.V22I01.036.

Haq, Mohd Anul, Abdul Khadar Jilani, and P. Prabu. 2022. "Deep Learning Based Modeling of Groundwater Storage Change." *Computers, Materials and Continua* 70(3):4599–4617. doi:10.32604/cmc.2022.020495.

Iqbal, Kashifa, Shamshad Ahmad, and Venkatesh Dutta. 2019. "Pollution Mapping in the Urban Segment of a Tropical River: Is Water Quality Index (WQI) Enough for a Nutrient-Polluted River?" *Applied Water Science* 9(8):1–16. doi:10.1007/s13201-019-1083-9.

Jiang, Simin, Jinhong Fan, Xuemin Xia, Xianwen Li, and Ruicheng Zhang. 2018. "An Effective Kalman Filter-Based Method for Groundwater Pollution Source Identification and Plume Morphology Characterization." *Water (Switzerland)* 10(8). doi:10.3390/w10081063.

Jigyasu, Dharmendra Kumar, Munendra Singh, Sandeep Singh, Satyendra Singh, and Indra Bir Singh. 2020. "Trace Element Mobility, Regional

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Significance and Global Implication of Gomati River Basin, Northern India." *SN Applied Sciences* 2(8):1–13. doi:10.1007/s42452-020-03204-0.

Kanti, Ajoy, Das Nandini, Gupta Tahir, Mahmood Binod, Chandra Tripathy, Rakhal Das, and Suman Das. 2024. "Assessing Anthropogenic Influences on the Water Quality of Gomati River Using an Innovative Weighted Fuzzy Soft Set Based Water Pollution Rating System." *Discover Water*. doi:10.1007/s43832-024-00136-3.

Kardos, Máté Krisztián, and Adrienne Clement. 2020. "Predicting Small Water Courses' Physico-Chemical Status from Watershed Characteristics with Two Multivariate Statistical Methods." *Open Geosciences* 12(1):71–84. doi:10.1515/geo-2020-0006.

Khan, Ramsha, and Abhishek Saxena. 2023. "Potentially Toxic Elements (PTEs) in Gomti-Ganga Alluvial Plain, Associated Human Health Risks Assessment and Potential Remediation Using Novel-Nanomaterials." *Environmental Monitoring and Assessment* 195(1). doi:10.1007/s10661-022-10562-2.

Khan, Ramsha, Abhishek Saxena, and Saurabh Shukla. 2020. "Evaluation of Heavy Metal Pollution for River Gomti, in Parts of Ganga Alluvial Plain, India." *SN Applied Sciences* 2(8):1–12. doi:10.1007/s42452-020-03233-9.

Khan, Ramsha, Abhishek Saxena, Saurabh Shukla, Pooja Goel, Prosun Bhattacharya, Peiyue Li, Esmat F. Ali, and Sabry M. Shaheen. 2022. "Appraisal of Water Quality and Ecological Sensitivity with Reference to Riverfront Development along the River Gomti, India." *Applied Water Science* 12(1):1–12. doi:10.1007/s13201-021-01560-9.

Khan, Ramsha, Abhishek Saxena, Saurabh Shukla, Selvam Sekar, and Pooja Goel. 2021. "Effect of COVID-19 Lockdown on the Water Quality Index of River Gomti, India, with Potential Hazard of Faecal-Oral Transmission." *Environmental Science and Pollution Research* 28(25):33021–29. doi:10.1007/s11356-021-13096-1.

Khullar, Sakshi, and Nanhey Singh. 2021. "Machine Learning Techniques in River Water Quality Modelling: A Research Travelogue." *Water Science and Technology: Water Supply* 21(1). doi:10.2166/ws.2020.277.

Kouadri, Saber, Ahmed Elbeltagi, Abu Reza Md Towfiqul Islam, and Samir Kateb. 2021. "Performance of Machine Learning Methods in Predicting Water Quality Index Based on Irregular Data Set: Application on Illizi Region (Algerian Southeast)." *Applied Water Science* 11(12):1–20. doi:10.1007/s13201-021-01528-9.

Krishnamoorthy, Loganathan, and Vignesh Rajkumar Lakshmanan. 2024. "Groundwater Quality Assessment Using Machine Learning Models: A Comprehensive Study on the Industrial Corridor of a Semi-Arid Region." *Environmental Science and Pollution Research* (July). doi:10.1007/s11356-024-34119-7.

Kumar, Pankaj. 2018. "Simulation of Gomti River (Lucknow City, India) Future Water Quality under Different Mitigation Strategies." *Heliyon* 4(12):e01074. doi:10.1016/j.heliyon.2018.e01074.

Kushwah, Vinod Kumar, Kunwar Raghvendra Singh, Nakul Gupta, Parveen Berwal, Faisal M. Alfaisal, Mohammad Amir Khan, Shamshad Alam, and Obaid Qamar. 2023. "Assessment of the Surface Water Quality of the Gomti River, India, Using Multivariate Statistical Methods." *Water (Switzerland)* 15(20). doi:10.3390/w15203575.

Liu, Xiang, and Zhiqiang John Zhai. 2009. "Prompt Tracking of Indoor Airborne Contaminant Source Location with Probability-Based Inverse Multi-Zone Modeling." *Building and Environment* 44(6):1135–43. doi:10.1016/j.buildenv.2008.08.004.

Mohinuddin, Sk, Soumita Sengupta, Biplab Sarkar, Ujwal Deep Saha, Aznarul Islam, Abu Reza Md Towfiqul Islam, Zakir Md Hossain, Sadik Mahammad, Taushik Ahamed, Raju Mondal, Wanchang Zhang, and Aimun Basra. 2023. "Assessing Lake Water Quality during COVID-19 Era Using Geospatial Techniques and Artificial Neural Network Model." *Environmental Science and Pollution Research* 30(24):65848–64. doi:10.1007/s11356-023-26878-6.

El Morabet, Rachida, Larbi Barhazi, Soufiane Bouhafa, Mohammed Abdullah Dahim, Roohul Abad Khan, and Nadeem A. Khan. 2023. "Geospatial Distribution and Machine Learning Algorithms for Assessing Water Quality in Surface Water Bodies of Morocco." *Scientific Reports* 13(1):1–15. doi:10.1038/s41598-023-47991-z.

Mukherjee, Sahana, Jayeeta Saha, Neha Sharma, Sovik Das, and Shiva Shankar Chaturvedi. 2024. "Water Quality Analysis, Treatment, and Economic Feasibility of Water Services of the Neora River in the Fringe Area of Neora-Valley National Park, India." *Water Supply* 24(8):2627–40. doi:10.2166/ws.2024.168.

Nandi, B. P., G. Singh, A. Jain, and D. K. Tayal. 2024. "Evolution of Neural Network to Deep Learning in Prediction of Air, Water Pollution and Its Indian Context." *International Journal of Environmental Science and Technology* 21(1):1021–36. doi:10.1007/s13762-023-04911-y.

Nayak, Anjali, Gagan Matta, and D. P. Uniyal. 2023. *Hydrochemical Characterization of Groundwater Quality Using Chemometric Analysis and Water Quality Indices in the Foothills of Himalayas*. Vol. 25. Springer Netherlands.

Paul, Dipak. 2017. "Research on Heavy Metal Pollution of River Ganga: A Review." *Annals of Agrarian Science* 15(2):278–86. doi:10.1016/j.aasci.2017.04.001.

Pimparkar, A. M., S. N. Patil, B. D. Patil, and A. K. Kadam. 2023. "Comparative Assessment of Wetland Water Quality from Rural and Urban Area of Aurangabad District, Maharashtra, India Using Water Quality Index." *HydroResearch* 6:269–78. doi:10.1016/j.hydres.2023.10.001.

Saalidong, Benjamin M., Simon Appah Aram, Samuel Otu, and Patrick Osei Lartey. 2022. "Examining the Dynamics of the Relationship between Water PH and Other Water Quality Parameters in Ground and Surface Water Systems." *PLoS ONE* 17(1 1):1–17. doi:10.1371/journal.pone.0262117.

Santy, Sneha, Pradeep Mujumdar, and Govindasamy Bala. 2020. "Potential Impacts of Climate and Land Use Change on the Water Quality of Ganga River around the Industrialized Kanpur Region." *Scientific Reports* 10(1):1–13. doi:10.1038/s41598-020-66171-x.

Saqib, Nazmu, Praveen Kumar Rai, Shruti Kanga, Deepak Kumar, Bojan Đurin, and Suraj Kumar Singh. 2023. "Assessment of Ground Water Quality of Lucknow City under GIS Framework Using Water Quality Index (WQI)." *Water (Switzerland)* 15(17). doi:10.3390/w15173048.

Sidek, L. M., H. A. Mohiyaden, M. Marufuzzaman, N. S. M. Noh, Salim Heddam, Mohammad Ehteram, Ozgur Kisi, and Saad Sh Sammen. 2024. "Developing an Ensembled Machine Learning Model for Predicting Water Quality Index in Johor River Basin." *Environmental Sciences Europe* 36(1). doi:10.1186/s12302-024-00897-7.

Singh, Aditya Pratap, Anshika Pandey, Aditya Kumar, Anju Chaurausiya, Rishabh Kashyap, Arstu Gautam, Mukul Saxena, and Hrishikesh Singh. 2023. "Water Quality Assessment of Gomti River by Using Modelling Technique: A Review." *Proceedings of the International Conference on Frontiers in Desalination, Energy, Environment and Material Sciences for Sustainable Development* (January):261–68. doi:10.21467/proceedings.161.30.

Singh, Kunwar P., Ankita Basant, Amrita Malik, and Gunja Jain. 2009. "Artificial Neural Network Modeling of the River Water Quality-A Case Study." *Ecological Modelling* 220(6):888–95. doi:10.1016/j.ecolmodel.2009.01.004.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 798-812

Singh, Kunwar P., Amrita Malik, and Sarita Sinha. 2005. "Water Quality Assessment and Apportionment of Pollution Sources of Gomti River (India) Using Multivariate Statistical Techniques - A Case Study." *Analytica Chimica Acta* 538(1–2):355–74. doi:10.1016/j.aca.2005.02.006.

Singh, Nidhi, and Smita Tung. 2025. "Assessment of Water Quality of Gomti River at Lucknow." *Air, Soil and Water Research* 18. doi:10.1177/11786221251328589.

Singh, Raj Mohan, and Bithin Datta. 2007. "Artificial Neural Network Modeling for Identification of Unknown Pollution Sources in Groundwater with Partially Missing Concentration Observation Data." *Water Resources Management* 21(3):557–72. doi:10.1007/s11269-006-9029-z.

Sohn, Insoo. 2021. "Deep Belief Network Based Intrusion Detection Techniques: A Survey." *Expert Systems with Applications* 167:114170. doi:10.1016/j.eswa.2020.114170.

Surya Prakash, Mishra. 2014. "Analysis of Water Quality of Gomti River At District Sultanpur (U.P.)." *International Journal of Engineering Science Invention Research & Development* I(Iii):113.

Syeed, M. M. Mahbubu., Md Shakhawat Hossain, Md Rajaul Karim, Mohammad Faisal Uddin, Mahady Hasan, and Razib Hayat Khan. 2023. "Surface Water Quality Profiling Using the Water Quality Index, Pollution Index and Statistical Methods: A Critical Review." *Environmental and Sustainability Indicators* 18(March):100247. doi:10.1016/j.indic.2023.100247.

Wang, Borui, Zhifang Tan, Wanbao Sheng, Zihao Liu, Xiaoqi Wu, Lu Ma, and Zhijun Li. 2024. "Identification of Groundwater Contamination Sources Based on a Deep Belief Neural Network." *Water (Switzerland)* 16(17). doi:10.3390/w16172449.

World Health Organization. (2008). *Guidelines for drinking-water quality* (3rd ed., Vol. 1). WHO Press. https://www.who.int/publications/i/item/9789241547611

Zhao, Ying, Ruizhuo Qu, Zhenxiang Xing, and Wenxi Lu. 2020. "Identifying Groundwater Contaminant Sources Based on a KELM Surrogate Model Together with Four Heuristic Optimization Algorithms." *Advances in Water Resources* 138(February):103540. doi:10.1016/j.advwatres.2020.103540.

Zheng, Hongmei, Shiwei Hou, Jing Liu, Yanna Xiong, and Yuxin Wang. 2024. "Advanced Machine Learning and Water Quality Index (WQI) Assessment: Evaluating Groundwater Quality at the Yopurga Landfill." *Water (Switzerland)* 16(12). doi:10.3390/w16121666.

List of Abbreviations

Abbreviation	Full Form
ANN	Artificial Neural Network
BOD	Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
DTR	Decision Tree Regression
DO	Dissolved Oxygen
EC	Electrical Conductivity
ML	Machine Learning
MLR	Multiple Linear Regression
NTU	Nephelometric Turbidity Unit
PCA	Principal Component Analysis
PI	Prediction Interval
R^2	Coefficient of Determination
RFR	Random Forest Regression
RMSE	Root Mean Square Error
SDGs	Sustainable Development Goals
SO_4^{2-}	Sulfate Ion
SVR	Support Vector Regression
TDS	Total Dissolved Solids
WQI	Water Quality Index
WWTP	Wastewater Treatment Plant