Environmental Science and Pollution Research (2025) 32:4670–4689
https://doi.org/10.1007/s11356-025-35999-z

RESEARCH ARTICLE

Integrating machine learning models for optimizing ecosystem health assessments through prediction of nitrate–N concentrations in the lower stretch of Ganga River, India

Basanta Kumar Das[1] · Sanatan Paul[1] · Biswajit Mandal[1] · Pranab Gogoi[1] · Liton Paul[1] · Ajoy Saha[1] · Canciyal Johnson[1] · Akankshya Das[1] · Archisman Ray[1] · Shreya Roy[1] · Shubhadeep Das Gupta[1]

Received: 3 September 2024 / Accepted: 21 January 2025 / Published online: 30 January 2025

Abstract

Nitrate, a highly reactive form of inorganic nitrogen, is commonly found in aquatic environments. Understanding the dynamics of nitrate–N concentration in rivers and its interactions with other water-quality parameters is crucial for effective freshwater ecosystem management. This study uses advanced machine learning models to analyse water quality parameters and predict nitrate–N concentrations in the lower stretch of the Ganga River from the observations of six annual periods (2017 to 2022). The parameters include water temperature, pH, specific conductivity (Sp_Con), dissolved oxygen (DO), nitrate–N, total phosphate (TP), turbidity, biochemical oxygen demand (BOD), silicate, total dissolved solids (TDS), and rainfall. The present study evaluated the predictive performance of five models—Multiple Polynomial Regression (MPR), Generalized Additive Models (GAMs), Decision Tree Regression, Random Forest (RF), and XGBoost (Extreme Gradient Boosting)—using RMSE, MAE, MAPE, NSE and R^2 metrics. XGBoost emerged as the top performer, with an RMSE of 0.024, MAE of 0.018, MAPE of 51.805, NSE of 0.855 and R^2 of 0.85, explaining 85% of the variance in nitrate–N concentrations. Random Forest also demonstrated strong predictive capability, with an RMSE of 0.028, MAE of 0.021, MAPE of 57.272, NSE of 0.804 and R^2 of 0.80. MPR effectively modelled non-linear relationships, explaining 75% of the variance, while Decision Tree Regression and GAMs were less effective, with R^2 values of 0.60 and 0.48, respectively. Variables (BOD, pH, Rainfall, water temperature, and total phosphate) were the best predictors of nitrate–N dynamics. Comparative analysis with previous studies confirmed the robustness of XGBoost and Random Forest in environmental data modelling. The findings highlight the importance of advanced machine learning models in accurately predicting water quality parameters and facilitating proactive management strategies.

Keywords Nitrate–N · Ganga River · GAMs · RF · XGboost · RMSE · MAE · NSE · Variable importance

Introduction

One of the most reactive forms of inorganic nitrogen found in aquatic environments is nitrate–N (Camargo et al. 2005), and it is essential to the nitrogen cycle, that sustains life on Earth. Nitrate–N in rivers comes from various sources, such as surface runoff, groundwater, atmospheric deposition, and the biological decomposition of organic materials in freshwater environments. The nitrogen cycle is a fundamental biogeochemical process that regulates the movement of nitrogen through plants, animals, bacteria, soil, water, and the atmosphere. This cycle involves critical steps including nitrogen fixation, nitrification, assimilation, and denitrification, interlinking the atmosphere, hydrosphere, and lithosphere. Nitrogen, an essential element of life, is a key component of amino acids, nucleic acids, and adenosine triphosphate (ATP), the primary energy molecule in all living organisms (Fenice 2021). The nitrogen cycle's components serve as electron acceptors for various bacteria, archaea, and eukaryotes, playing crucial roles in ecosystems and industrial applications, such as agriculture and wastewater treatment, where

Responsible Editor: Marcus Schulz

✉ Basanta Kumar Das
basantakumard@gmail.com

[1] ICAR-Central Inland Fisheries Research Institute, Barrackpore, Kolkata 700120, West Bengal, India

nitrogen metabolism impacts greenhouse gas emissions and public health (Kuypers et al. 2018).

Besides water quality parameters, anthropogenic effects may alter the ecosystem's balance. Moreover, human activities related to agriculture, industry, and urbanization increasingly contribute to nitrate–N levels (Korostynska et al. 2012). In particular, fossil fuel combustion contributes significantly to atmospheric nitrate–N deposition (Boumans et al. 2004). In its bioavailable form, nitrate–N is used by riverine organisms such as algae, macrophytes, and certain bacteria for growth and metabolism, forming the foundational elements of aquatic food webs (Camargo et al. 2005). However, excessive nitrate–N levels can lead to eutrophication, causing issues like reduced light penetration and lower dissolved oxygen concentrations (Bricker et al. 2008). These conditions can harm aquatic life, including invertebrates and fish (Moore and Bringolf 2018). Therefore, understanding the dynamics of nitrate–N concentration in rivers and its interactions with other water-quality parameters is vital for effectively managing freshwater ecosystems.

Motivation

Human activities have significantly intensified the global nitrogen cycle, resulting in environmental disturbances such as biodiversity loss, climate change, and eutrophication in aquatic ecosystems (Howarth et al. 2012). Nitrate, a critical pollutant, remains challenging to analyze accurately, with limited research available on its contamination in river systems. Globally, nitrogen inputs to rivers have increased substantially (Green et al. 2004), with riverine discharges being key pathways for nitrogen compounds such as ammonium, nitrite and nitrate, and into coastal waters (Gao et al. 2022). The rise in human settlements and agricultural practices near rivers exacerbates nitrogen loading, leading to eutrophication and ecosystem disruptions (Murphy 2022; Li et al. 2023). The Ganga River, particularly its lower stretch, is a vital yet vulnerable resource, heavily impacted by pollution pressures from high human activity and industrialization (Kumar et al. 2024). Predicting nitrate levels is critical for monitoring water quality, especially since even minor increases in nitrate can have significant ecological impacts (Moore and Bringolf 2018). However, despite evident risks, limited studies have analyzed nitrate (Canter 2019; Moore and Bringolf 2018) levels or developed predictive frameworks to address this concern. This study aims to bridge the gap by developing a predictive model for nitrate contamination. Such a model provides an early warning system for water quality management, helping to prevent nitrate levels from exceeding safe thresholds.

Prior research on the use of machine learning and deep learning algorithms to predict nitrate–N and water quality metrics

Predicting nitrate–N concentrations in aquatic environments is critical for managing water quality and preventing eutrophication, which can lead to harmful algal blooms (Bricker et al. 2008). Traditional methods of nitrate–N removal, such as reverse osmosis and ion exchange, have limitations in efficiency and cost. Microbial bioremediation offers a promising alternative, yet its complexity necessitates advanced predictive tools (Canter 2019). Several studies have focused on predicting water quality parameters using machine learning and deep learning algorithms, in addition to nitrate–N prediction. Some of the recent studies are highlighted in this section. Artificial neural networks (ANN), support vector regression (SVR), and long short-term memory recurrent neural networks (LSTM-RNN) are examples of recent developments in machine learning (ML) and deep learning techniques that have been used to forecast turbidity in Hong Kong's marine environment. The most successful model for this purpose was LSTM-RNN, which performed better than SVR and ANN, with an accuracy of 88.45%, compared to 81% and 85.6% for SVR and ANN, respectively (Kumar et al. 2022).

In environmental modelling, nitrate–N and phosphorus concentrations have been estimated across five watersheds—the Cuyahoga, Grand, Maumee, Raisin, and Sandusky rivers—in Ohio and Michigan using multiple linear regression (MLR), k-nearest neighbours (kNN), regression tree (RT), random forest (RF), artificial neural network (ANN), support vector machine (SVM), and Gaussian process regression (GPR) (Bhattara et al. 2021). A similar study was carried out by Jaddi et al. (2024) using artificial neural networks (ANN), support vector regression (SVR), and multiple linear regression (MLR) to estimate nitrate–N concentrations in groundwater at the Saiss aquifer in Northern Morocco. They employed the following six water quality parameters as predictors: SO_4^{2-}, HCO_3^-, EC, Cl^-, Ca^{2+}, and Na^+. With a mean square error (MSE) of 757.345 and an R^2 of 0.523, MLR yielded low accuracy. Nonetheless, both ANN and SVR models greatly increased prediction accuracy, achieving R^2 values of 0.836 with an MSE of 0.023 and, after additional optimization, $R^2=0.902$ and MSE $=4.364$. Similarly, Nair and Vijaya (2022) used predictive models such as random forest, MLP regressor, support vector regressor, and linear regression; classification methods included SVM, naïve Bayes, decision trees, and MLP classifiers for accurate prediction of water quality index from Bhavani River, Tamil Nadu. Singh et al. (2024) also studied the groundwater suitability for sustainable irrigation using indexical, statistical, and machine learning approaches. Several other studies in abroad such as Egbueri et al. (2024) have been well explained on the use of a machine learning approach to assess the quality of groundwater in the Volta basin and groundwater pollution and health risks in eastern Ghana, with a particular emphasis on fluoride (F^-) and nitrate (NO_3^-) contamination, respectively. Agbasi and Egbueri (2024) detected seven input variable groupings from their examination of 139 models, with an emphasis on metal, chemical, and physical factors in their study for identifying potentially toxic elements (PTEs) in water resources using MLP-NN (multilayer perceptron neural network), RBF-NN (radial basis function neural network), and ANFIS (adaptive neuro-fuzzy inference system). Despite MLP-NN being the most popular algorithm, ANFIS outperformed RBF-NN and MLP-NN according to performance rankings based on correlation and determination coefficients (R, R^2).

Ensemble methodologies like Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are extensively employed in water science but have limited application in predicting nitrate concentrations in surface and groundwater (Mahboobi et al. 2023). Research focusing specifically on the Ganga River Basin is minimal. RF is preferred for its ability to capture non-linear dependencies, computational efficiency, smooth parameterizations, and use of variable importance metrics (Elavarasan and Vincent 2021). Numerous studies have applied RF to predict nitrate patterns in groundwater (Abba et al. 2024; Mahlknecht et al. 2023). Conversely, XGBoost improves model robustness and execution speed through regularization terms, column sampling, and optimized decision tree split points (Ransom et al. 2022). In water quality research, XGBoost has demonstrated superior performance compared to algorithms like LogiBoost, RF, AdaBoost, and support vector machines (Shams et al. 2024). Critical factors influencing nitrate prediction include nitrogen, ammonium, phosphate, pH, rainfall, turbidity, water temperature, dissolved oxygen, and biological oxygen demand (Zhang et al. 2022).

Therefore, the present study aims to develop, validate, and compare machine learning prediction models to optimistic nitrogen management strategies and promote environmental sustainability in the lower stretch of the Ganga River. This study focuses on analyzing water quality parameters from nine locations i.e., Buxar, Patna, Bhagalpur, Farakka, Jangipur, Berhampore, Balagarh, Tribeni, and Godakhali, collected from 2016 to 2022. The primary objective is to predict the concentration of nitrate–N, integrating machine learning models such as Multiple polynomial regression (MPR), Generalized additive models (GAMs), Decision tree regression (DTR), Random Forest (RF) and XGboost. These methods allow for capturing complex, non-linear relationships between the independent variables such as total phosphate, pH, rainfall, turbidity, water temperature, DO, BOD, TDS and silicate, providing a comprehensive understanding of the factors affecting nitrate–N levels.

Materials and methods

Study area

The present study was conducted in the lower stretch of the Ganga River Basin, traversing the Indian states of Bihar and West Bengal, to predict nitrate–N contents in the river water. Despite being revered as a "Holy River," the pollution level in the Ganga River is gradually increasing due to exposure to urban and industrial waste, atmospheric deposition, agricultural runoff, and religious activities. The Indian government has prioritized the rejuvenation of the Ganga due to its critical importance. For this study, 9 (nine) sites along the Ganga River, covering approximately 1,208 km from Buxar to Godakhali, were selected. These sampling stations were strategically positioned to represent the river's middle and lower segments. Data were collected from locations viz Buxar (25° 34′ 26.616" N, 83° 57′ 55.296" E), Patna (25° 37′ 4.116" N, 85° 11′ 56.652" E), Bhagalpur (25° 19′ 1.164" N, 86° 59′ 50.712" E), Farakka (24° 48′ 29.88" N, 87° 55′ 21.72" E), Jangipur (24° 27′ 0.72" N, 88° 6′ 25.2" E), Berhampore (24° 5′ 39.876" N, 88° 14′ 39.948" E), Balagarh (23° 7′ 45.408" N, 88° 27′ 43.488" E), Tribeni (22° 59′ 17.016" N, 88° 24′ 19.836" E), Godakhali (22° 23′ 58.272" N, 88° 7′ 46.452" E). Water quality parameters recorded include water temperature (°C), pH, dissolved oxygen (mg/L), nitrate–N (mg/L), total phosphate (mg/L), turbidity (NTU), biochemical oxygen demand (mg/L), specific conductivity (mS/cm), silicate (mg/L), total dissolved solids (mg/L) and rainfall (mm). Hydrologically, the stretch from Buxar to Godakhali is categorized as freshwater. The lower stretch of the Ganga River, extending from Buxar to Godakhali, is marked by diverse physiographic patterns, topographical features, and seasonal variations influenced by climate and geological units. This area includes a freshwater riverine section with a sandy clay bed in a climate that sees monsoonal rainfall patterns and seasonal fluctuations in water levels. This is predominantly a freshwater zone and is unaffected by tidal influences. Moving downstream from Balagarh to Godakhali, the topography shifts to plains dominated by alluvial clays, forming an estuarine zone with geological deposits regularly reshaped by transported sediments. This section is characterized by medium to high tidal effects and a rising salinity regime, especially in locations like Godakhali, where the influence of the Bay of Bengal is more pronounced.

Tropical monsoon climate with hot summers, a distinct rainy season, and mild winters is experienced in Bihar, India. Temperatures throughout the summer (April to June) can reach 30°C to 45°C, and heavy humidity is frequently present. Arriving in June, the southwest monsoon brings heavy rainfall, with the wettest months being July and August.

Rainfall averages vary by location, ranging from 800 mm to 1,500 mm annually (Attri and Tyagi 2010). West Bengal's tropical monsoon climate differs significantly due to its diverse geography, from the Bay of Bengal coast to the Himalayan foothills. Summers (March to June) are hot and humid, with temperatures around 30–40°C and humidity above 70%, especially near the coast. The monsoon season (June to September) is crucial for agriculture, with heavy rainfall often exceeding 1500 mm, though it can bring cyclones and flooding, especially in coastal regions. Winters (December to February) are mild in the plains (10–20°C) and colder in the northern hills, supporting rich biodiversity and varied farming practices (Patra and Bhaskaran 2016). The locations of these sampling sites are depicted in Fig. 1.

Sampling strategies

Water samples were collected from nine designated sites during three seasonal periods, i.e. pre-monsoon (Feb-May), monsoon (June-Sep), and post-monsoon (Oct-Jan). Surface water samples were collected in triplicate using clean, pre-rinsed polyethylene bottles. To ensure consistency, Tarsons® HDPE bottles were thoroughly rinsed with river water before each collecting water sample. The sampling was carried out from 2017 to 2022.

Field parameters

Field parameters were measured in situ, including water temperature, pH, specific conductivity, DO, nitrate–N, TP, turbidity, BOD, silicate, TDS, and rainfall. Water temperature, pH, specific conductivity, and TDS were measured using the AQUAREAD 7000® multi-parameter water quality analysis instrument. Dissolved oxygen and BOD were determined using the titrimetric method (APHA 2012). The collected samples were immediately stored in ice-cooled containers to prevent any changes in water quality during transit. Samples designated for further laboratory analysis were kept in iceboxes until they could be processed. In the laboratory, filtered water samples were analyzed for nutrients using UV–Vis spectrophotometry. Precisely, silicate (SiO_4-Si), TP, and nitrate–N (NO_3-N) were measured following standard protocols (APHA 2012). All sampling, storage, preservation, and analysis procedures strictly adhered to the guidelines provided by APHA (2012).

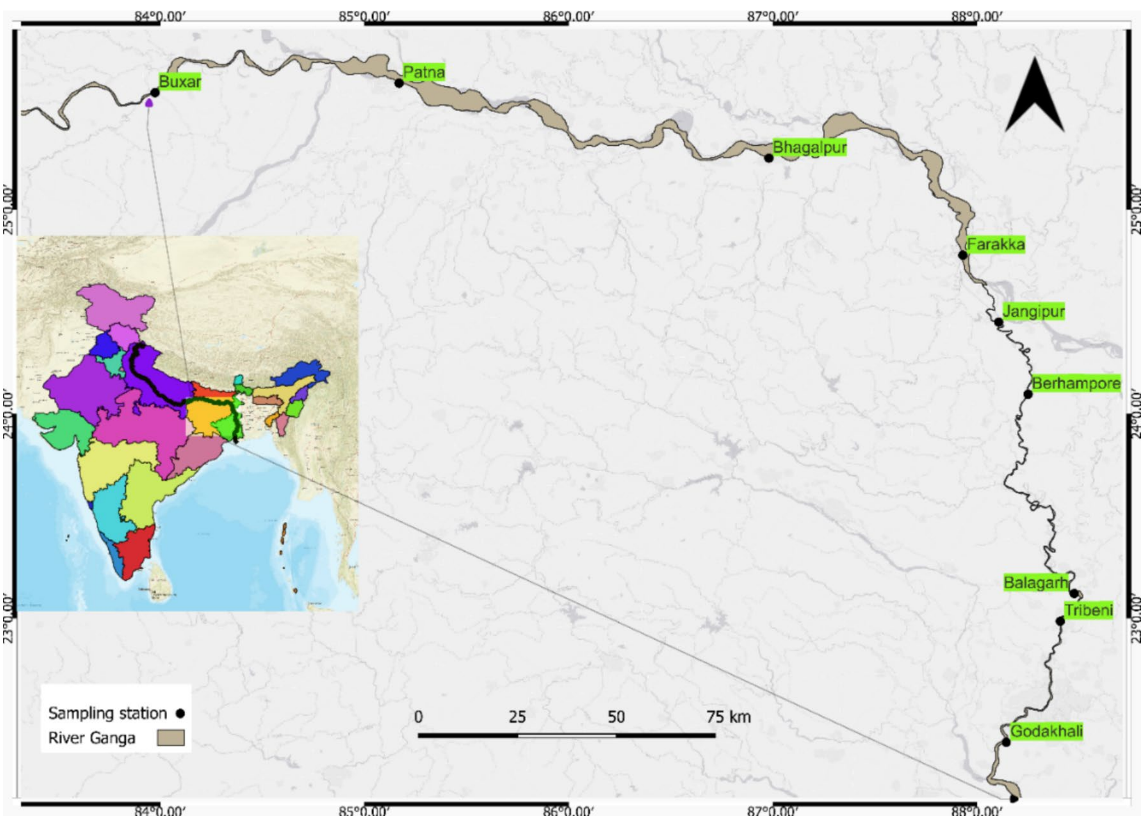

Fig. 1 Study area map showing the sampling sites in the Lower Ganga River

Statistical analysis

Descriptive statistics

Descriptive statistics, including means and standard deviations (SD), were computed for all water quality variables to summarize the data distribution and identify underlying patterns. These measures provide a comprehensive overview of the central tendency and dispersion of the water quality parameters across different seasons and geographical locations. In addition, to understand the significant variations of water parameters between seasons and stations Permutational MANOVA was performed.

Model derivation

Multiple Polynomial Regression (MPR)

Multiple Polynomial Regression models the relationship between a dependent variable or response variable and multiple independent variables or explanatory variables using polynomial functions (Belany et al. 2024). This method captures non-linear relationships by incorporating polynomial terms of different orders (Ghanizadeh et al. 2024). The general form of the model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \cdots + \beta_{nn} X_n^2 + \cdots + \beta_{12} X_1 X_2 + \varepsilon \tag{1}$$

where, Y is the dependent variable, $X_1, X_2, \ldots, X_n$ are the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and ε is the error term (Aslan 2024). Using this method, the model may consider intricate relationships and higher-order effects amongst the predictors, giving a more accurate depiction of the fundamental mechanisms influencing water quality (Visser et al. 2024).

Generalized Additive Models (GAMs)

To further explore the non-linear relationships between the water quality variables and the response variable, Generalized Additive Models (GAMs) were employed. GAMs extend the Generalized Linear Models (GLMs) by consolidating smooth functions of the predictors, thereby allowing for flexible modelling of non-linear patterns (Lai et al. 2024). The smooth functions are typically represented using splines or other basis functions, and regularizations techniques are applied to control the smoothness and stop overfitting (Siems et al. 2024). The GAMs can be expressed as:

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) + \varepsilon \tag{2}$$

The model connects a single response variable, Y, to various predictor variables, $X_{i.}$ A distribution (such as a standard, binomial, or Poisson distribution) is defined from the exponential family for the response variable Y. In addition, a link function g (such as the identity or log functions) is used to link the predictor variables and the predicted value of Y to a model structure like this: $f_i(x_i)$ denotes the smooth function of the ith predictor. This model identifies intricate non-linear associations in the data, enhancing the interpretability and predictive accuracy of the analysis.

Decision Tree Regression (DTR)

The Decision Tree algorithm is a popular and intuitive machine-learning method used for both classification and regression tasks (Modhugu and Ponnusamy 2024). It works by splitting the dataset into subsets based on the value of input features, creating a tree-like model of decisions and their possible consequences (Malhotra et al. 2024) (Fig. 2). This study focuses on decision trees for regression. The goal of each subset is to maximise the homogeneity of the target variable. To split data, the optimal feature is chosen using metrics like mean squared error (MSE), and the splitting procedure is used recursively until a stopping requirement is satisfied (Parhi and Patro 2024).

Random Forest (RF)

The Random Forest algorithm is a strong ensemble technique effective for classification and regression applications. It works by building many decision trees throughout the training process, from which it produces a class that is either the mean prediction (regression) or the mode of the classes (classification) of the individual trees (Salman et al. 2024; Manzali et al. 2024) (Fig. 3). Every tree in the ensemble is constructed using a random subset of the training set, and a random subset of features is considered at every split in the tree (Sun et al. 2024; Khalifa et al. 2024). The model performs robustly while being less prone to overfitting because of the randomness's role in decorrelation of the trees (Bouke et al. 2024). Moreover, they provide measures of feature importance, aiding in understanding the contribution of each feature to the model's predictions (Banik et al. 2024).

XG boost (Extreme Gradient Boosting)

For further improvement in prediction accuracy and computational efficiency, XGBoost (Extreme Gradient Boosting) was employed. XGBoost is a scalable and optimised version of gradient boosting machines (Özdemir et al. 2024). It constructs an ensemble of weak learners, typically decision trees, and optimizes them using gradient descent combined with regularization techniques to prevent overfitting (Choudhury

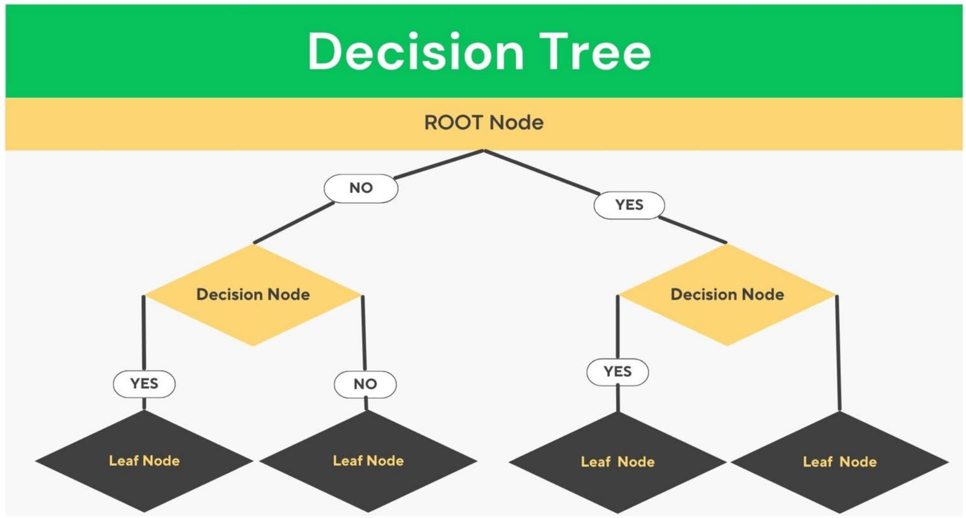

Fig. 2 Flow Chart of Decision Tree algorithm

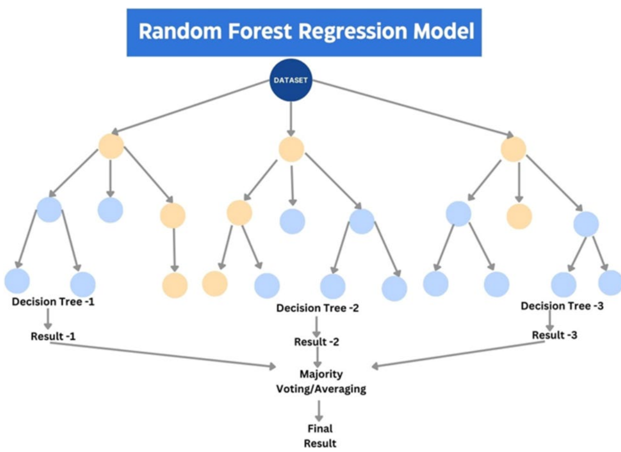

Fig. 3 Flow Chart of Random Forest algorithm

et al. 2024). XGBoost incorporates advanced features such as parallel computation, tree pruning, and handling of missing values, outperforming traditional gradient-boosting algorithms in speed and performance (Shaik et al. 2024). During training, hyperparameter tuning and early stopping criteria were used to enhance further the model's predictive accuracy (Ogunpola et al. 2024).

Model evaluation

The models were evaluated using several performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Nash–Sutcliffe Efficiency (NSE), and R-squared (R^2). RMSE quantifies the average magnitude of prediction errors, while MAE represents the average absolute difference between predicted and observed values. MAPE provides the mean percentage error, offering insight into prediction accuracy as a percentage. R^2 indicates the proportion of variance in the dependent variable that is explained by the independent variables (Heddam et al. 2023; Wang et al. 2023). NSE assesses model performance by comparing the variance of residuals to the variance of the observed data, offering a relative measure of prediction accuracy (Duc and Sawada 2023; Lamontagne et al. 2020).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{4}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{5}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value and n is the number of sample sizes.

The final model selection was based on a comprehensive assessment of these criteria, ensuring that the chosen approach provided accurate predictions and offered insights into the underlying relationships between the water quality variables. By integrating multiple modelling techniques, the study achieved a robust and nuanced understanding of the factors influencing water quality, paving the way for more effective management and policy decisions. The Decision Tree Regression, Random Forest (RF), XGBoost algorithms, and cross-validation (CV) were run using the Python ML library packages scikit-learn and XGBoost. A k-fold CV with ten repetitions was used to validate each model. A grid search for model performance optimization was conducted utilizing CV to execute Decision Tree Regression, RF, and XGBoost analysis. Using the 'gam' package and the 'mgcv' package, the GAM model analysis was carried out using the R statistical software.

Results

Assessment of physiochemical drivers on the regulation of River Ganga Health Ecosystem

The water quality metrics at nine separate places along the Ganga River (Buxar, Patna, Bhagalpur, Farakka, Jangipur, Berhampore, Balagarh, Tribeni, and Godakhali) and during three distinct seasonal periods (pre-monsoon, monsoon, and post-monsoon) are presented in detail in Table 1 and Fig. 4. The average water temperature usually varies between 27.55 ± 3.03°C (S4) to 29.09 ± 3.06°C (S7) across the sites. During the monsoon season, the average temperature dropped significantly (18.15 ± 15.58°C), likely due to increased rainfall and reduced solar radiation (Wang et al. 2024). The mean pH levels across the sites generally range from 7.87 ± 0.41(S9) to 8.26 ± 0.44(S6). pH levels dropped significantly during the monsoon season (4.59 ± 3.94), suggesting increased acidity due to runoff and precipitation dilution (Yotsuyanagi, et al. 2024). Dissolved oxygen (DO) levels across the sites average from 6.10 ± 1.11 mg/L at S3 to 6.90 ± 0.10 mg/L at S4. In contrast, DO levels dropped significantly during the monsoon season to 3.50 ± 3.06 mg/L, likely due to higher temperatures and increased organic waste decomposition (Singh 2024). The nitrate–N concentrations vary between 0.05 ± 0.05 mg/L(S2) to 0.11 ± 0.07 mg/L(S3) across the sites. In the pre-monsoon season, the nitrate–N level was 0.08 ± 0.06 mg/L. However, during the monsoon season, it dropped to its lowest at 0.03 ± 0.05 mg/L, likely due to rainfall-induced dilution (Dupas et al. 2024). The lowest total phosphate (TP) levels were found at S2 and S8, whereas the highest were at S4. Turbidity, which indicates water clarity, shows considerable variation across sites. The highest turbidity level of 27.58 ± 13.99 NTU was recorded at S7, while the lowest level of 13.02 ± 13.49 NTU was observed at S9. Turbidity tends to be higher in the pre-monsoon and post-monsoon seasons than in the monsoon season, likely due to sediment resuspension and reduced flow rates outside the monsoon period (Chakraborty et al. 2024). Biochemical oxygen demand (BOD), an indicator of organic pollution, ranges

Table 1 Variation of water variables in the study sites of Ganga River ($n = 110$; showing mean values and standard deviation)

Variables	S1	S2	S3	S4	S5	S6	S7	S8	S9	PRM	MON	POM
Temp (°C)	28.85 ± 3.95	28.99 ± 2.96	28.19 ± 3.47	27.55 ± 3.03	28.50 ± 3.13	28.23 ± 2.94	29.09 ± 3.06	28.38 ± 3.38	29.01 ± 2.87	28.09 ± 3.01	18.15 ± 15.58	21.07 ± 11.54
pH	7.9 ± 0.59	8.03 ± 0.38	8.14 ± 0.35	8.258 ± 0.44	8.20 ± 0.35	8.26 ± 0.32	8.13 ± 0.47	8.06 ± 0.42	7.87 ± 0.41	8.1 ± 0.42	4.59 ± 3.94	6.45 ± 3.45
BOD (mg/L)	1.45 ± 0.62	1.42 ± 0.72	1.17 ± 0.62	1.061 ± 0.64	1.22 ± 0.88	1.13 ± 0.71	1.25 ± 0.74	1.60 ± 0.92	1.04 ± 0.65	1.13 ± 0.66	0.90 ± 0.88	0.96 ± 0.94
DO (mg/L)	6.60 ± 0.97	6.32 ± 0.73	6.10 ± 1.11	6.896 ± 0.10	6.61 ± 1.36	6.66 ± 1.00	6.54 ± 1.19	6.28 ± 1.06	6.45 ± 1.48	6.54 ± 0.89	3.50 ± 3.06	5.34 ± 3.09
Sp_Con (mS/cm)	0.41 ± 0.11	0.32 ± 0.08	0.36 ± 0.11	0.30 ± 0.09	0.30 ± 0.11	0.31 ± 0.09	0.29 ± 0.07	0.30 ± 0.07	0.43 ± 0.15	0.38 ± 0.10	0.14 ± 0.13	0.25 ± 0.15
Turbidiy (NTU)	16.92 ± 7.06	27.58 ± 13.99	20.05 ± 10.12	13.027 ± 13.49	13.89 ± 4.51	17.55 ± 7.37	25.88 ± 9.86	22.83 ± 13.20	25.08 ± 14.95	17.75 ± 10.61	11.28 ± 12.78	17.64 ± 13.04
TDS (g/l)	0.23 ± 0.09	0.20 ± 0.07	0.22 ± 0.10	0.200 ± 0.08	0.19 ± 0.04	0.21 ± 0.09	0.19 ± 0.07	0.17 ± 0.05	0.25 ± 0.08	0.22 ± 0.07	0.11 ± 0.11	0.15 ± 0.11
Nitrate (mg/L)	0.09 ± 0.06	0.11 ± 0.07	0.11 ± 0.07	0.08 ± 0.07	0.06 ± 0.05	0.05 ± 0.05	0.05 ± 0.05	0.06 ± 0.05	0.10 ± 0.08	0.08 ± 0.06	0.03 ± 0.05	0.07 ± 0.07
TP (mg/L)	0.09 ± 0.05	0.08 ± 0.04	0.11 ± 0.06	0.11 ± 0.08	0.09 ± 0.06	0.10 ± 0.07	0.09 ± 0.06	0.08 ± 0.04	0.09 ± 0.08	0.10 ± 0.06	0.04 ± 0.05	0.08 ± 0.07
Silicate (mg/L)	7.8 ± 3.03	7.91 ± 2.65	7.63 ± 2.12	7.807 ± 2.55	7.49 ± 2.85	8.10 ± 3.38	7.83 ± 2.60	8.44 ± 2.45	8.57 ± 3.47	6.85 ± 2.47	5.64 ± 5.28	6.26 ± 3.87
Rainfall (mm)	47.38 ± 66.75	67.65 ± 80.97	47.78 ± 54.91	67.67 ± 80.39	70.82 ± 81.52	71.35 ± 76.94	99.02 ± 77.56	103.45 ± 85.99	76.40 ± 75.74	54.28 ± 61.90	88.89 ± 92.24	29.14 ± 49.16

PRM: pre-monsoon; MON: monsoon; POM: Post-monsoon

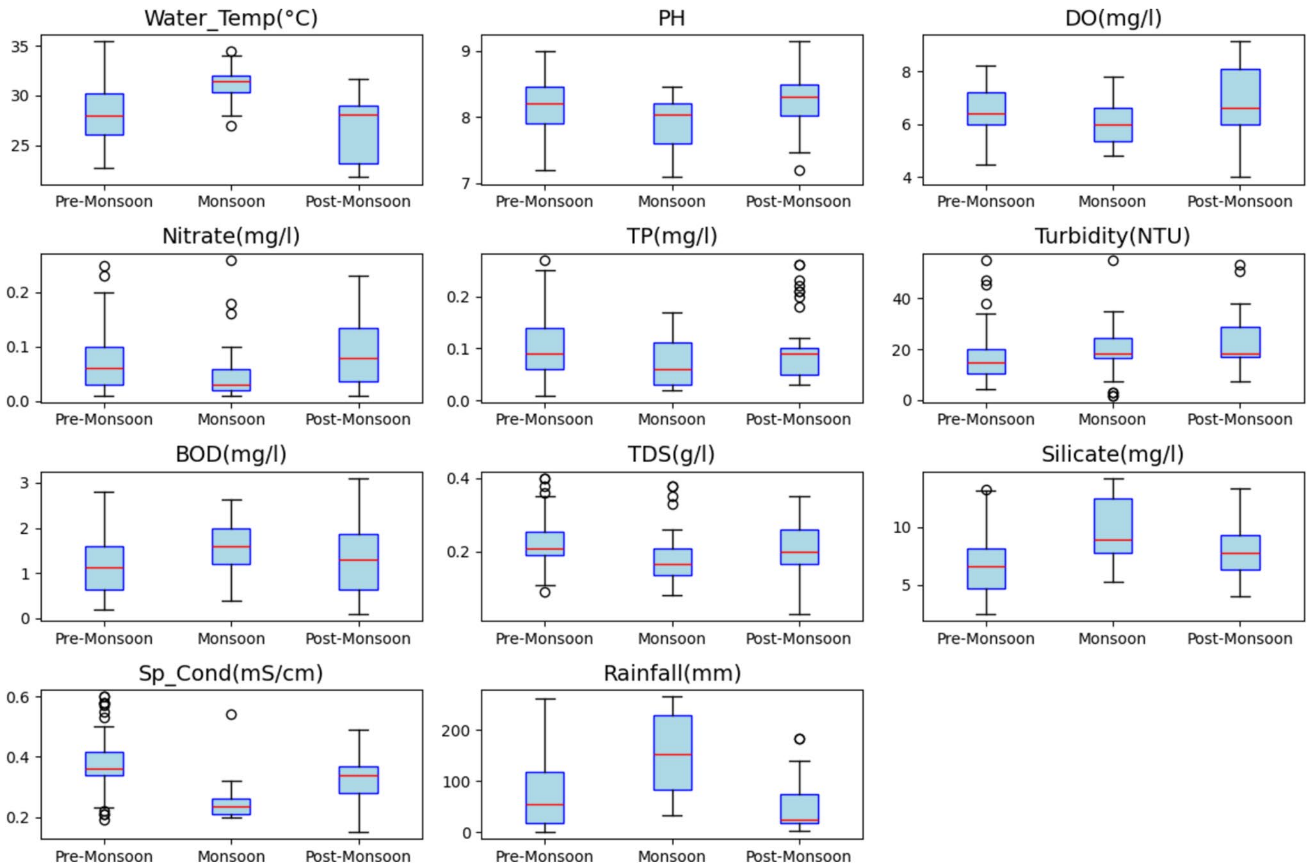

Fig. 4 Box plot showing the seasonal changes (pre-monsoon, monsoon, and post-monsoon) in water variables and rainfall in the lower stretch of the Ganga River

from 1.04 ± 0.65 (S9) to 1.60 ± 0.92 (S8) mg/L across the sites. The lowest (1.13 ± 0.66 mg/L) BOD was recorded during the monsoon season among the seasons. Total dissolved solids (TDS), indicating the concentration of dissolved substances in water, remain low, ranging from 0.17 ± 0.05(S8) to 0.25 ± 0.08(S9) g/L across the sites. The lowest values are observed during the monsoon season (0.11 ± 0.11 g/L), consistent with the dilution effect of increased rainfall (Patra et al. 2024). Silicate levels, important for diatom growth, range from 7.63 ± 2.12(S3) to 8.57 ± 3.47(S9) mg/L across the sites but drop to 5.64 ± 5.28 during the monsoon, possibly due to lower inputs from terrestrial sources (Senarathne et al. 2024). Specific conductivity (Sp_Con), which measures the water's ability to conduct electrical current, varies from 0.29 ± 0.07(S7) to 0.43 ± 0.159(S9) mS/cm across the sites. It drops significantly during the monsoon season (0.14 ± 0.13 mS/cm) due to the dilution effect (Naseri Boroujeni et al. 2024). Rainfall varies greatly across locations and seasons, peaking during the monsoon season at 88.89 ± 92.24 mm. The highest rainfall is recorded at S8 (Tribeni) with 103.45 ± 85.99 mm. This variability significantly influences water quality parameters, underscoring the impact of seasonal changes on river health (Bhatt et al. 2024). PERMANOVA analysis depicted a significant variation in water variables between seasons ($F = 12.019$; $p = 0.001$; $F = 1.3899$; $p = 0.033$) and stations (Table 2). In the PERMANOVA table, the degree of freedom accounted for seasons (df: 3–1 = 2) and stations (df: 9–1 = 8). Here, the sum of squares (SS) measures the overall variability. Higher SS values signify a greater contribution to the total variation. Most variability is explained by seasons (SS = 223.68), but stations (SS = 77.598) also play a role. A high mean square for seasons (MS = 111.84) indicates a substantial seasonal effect, but a smaller effect is seen for stations (MS = 12.933). Seasons and stations cannot account for the background variability captured by residuals (MS = 9.3051). The detected differences' statistical significance is shown by the P-value. It is confirmed that there are strong seasonal influences on water variables by the highly significant value for seasons ($P = 0.001, < 0.01$). The P-value for stations ($P = 0.033, < 0.05$) shows moderate but considerable spatial diversity among the seven sites. It highlighted the significant impact of geographical (station-based) and temporal (seasonal) factors on the variability of water variables.

Table 2 PERMANOVA showing the significant variations of water variables in seasons and stations

Source	df	SS	MS	Pseudo- F	P	perms
Seasons	2	223.68	111.84	12.019	**0.001**	997
Stations	8	77.598	12.933	1.3899	**0.033**	998
Res	105	977.03	9.3051			

Multiple Polynomial Regressions (MPR)

The MPR (Multivariate Polynomial Regression) technique predicts nitrate–N concentrations by modeling complex relationships among environmental variables, effectively capturing nonlinear interactions in riverine ecosystems. To fit an MPR model, the optimal degree by fitting multiple models with varying degrees of complexity and comparing their performance was determined.

Various degrees of MPR models were assessed:

- Degree 1: RMSE = 0.056, MAE = 0.043, MAPE = 104.625, NSE = 0.222, $R.^2 = 0.22$
- Degree 2: RMSE = 0.039, MAE = 0.029, MAPE = 76.168, NSE = 0.621, $R.^2 = 0.62$
- Degree 3: RMSE = 0.030, MAE = 0.023, MAPE = 58.313, NSE = 0.754, $R.^2 = 0.75$

The degree 3 MPR model performed best, achieving an R^2 of 0.75, RMSE of 0.030, MAE of 0.023, MAPE of 58.313, and NSE of 0.754. The low RMSE, MAE, and MAPE values demonstrated high accuracy and minimal prediction error. To avoid overfitting, the MPR model is limited to a maximum degree of 3. The accompanying plot in Fig. 5 shows the actual values, predicted values, and MPR curves for model degrees 1–3, highlighting improved accuracy with higher degrees.

Generalized Additive Models (GAMs)

Generalized Additive Models (GAMs) were employed to explore the relationships between nitrate–N concentration and various covariates. The smooth curves generated by these regressions revealed distinct patterns of association for each covariate (Dovers et al. 2024). These coefficients indicate the strength and direction of the relationship between the different water quality parameters and the nitrate–N concentration (McDowell et al. 2024).

In the present model, all the water quality parameters have positive coefficients signifies water quality parameters had a positive association with nitrate–N concentration. GAMs provide a measure of deviance explained by the model, explaining 48% of deviance in the data. This indicates that the model does not fit the observed data well. In Generalized Additive Models (GAMs) (Fig. 6), the bold line shows the relationship between a predictor and nitrate–N concentration, highlighting non-linear trends. The dashed lines around it represent the 95% confidence interval—narrow bands indicate higher confidence, while wider bands suggest more uncertainty. Each sub-plot shows the effect of one water quality parameter (e.g., pH, DO, temperature, turbidity etc.) on nitrate–N, accounting for other variables. The y-axis represents nitrate–N concentration, and the x-axis shows each predictor. Figure 6 shows that water temperature and nitrate–N concentration are linearly related, with narrower confidence intervals except at extreme temperatures. pH also has a nearly linear relationship with nitrate–N slightly increasing at higher values with similar confidence intervals. Figure 6 reveals a non-linear relationship between DO and nitrate–N where it increases up to 6.5 mg/L decreases until 8 mg/L and then rises again with wider confidence intervals at extremes. TP shows an inverse relationship with nitrate–N while turbidity increases nitrate–N up to 20 NTU before declining both with wider intervals. BOD inversely affects nitrate–N with narrower confidence intervals. Rainfall, silicate, and TDS have minimal effects but more precise estimates, while specific conductivity shows an oscillatory pattern with higher uncertainty.

Decision Tree Regression (DTR)

This study focuses on a Decision Tree regression model, which is flexible enough to handle both linear and non-linear data. The tree's maximum depth is set to 4, meaning it can grow up to four levels. Deeper trees have more splits, making the decisions more complex, while shallower trees are simpler. The minimum sample split is set to 3, so if a node has fewer than three samples, it won't split further. The decision tree flowchart (Fig. 7) shows the threshold values at each node, along with the sample size and Mean Squared Error (MSE), which measures how far off the predictions are from the actual values. The tree splits at each node based on specific conditions, continuing recursively to make predictions (Schiller et al. 2024; Negru 2024; Lee et al. 2024).

The root node splits the data at a BOD threshold of 0.49, dividing 111 samples with an average value of 0.075 and minimising the squared error to 0.004 (Fig. 8). The left subtree splits further based on Total Phosphate ($TP \leq 0.095$) and pH values (≤ 7.675, ≤ 8.15), resulting in leaf nodes with squared errors of 0.001 and 0.002, and

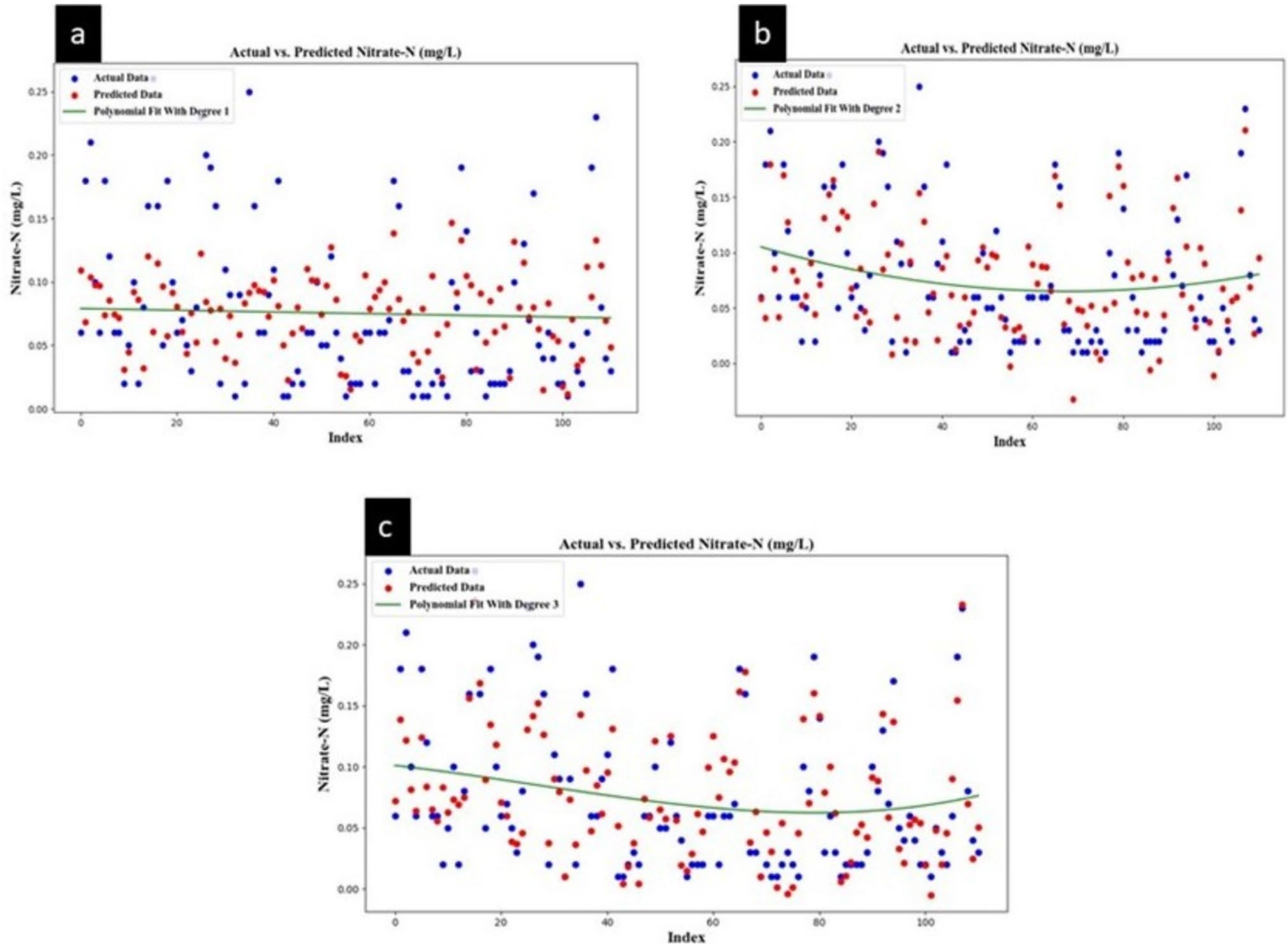

Fig. 5 Polynomial regression models (degrees 1, 2, and 3) fitted to nitrate–N concentration data, showing their fit to the observed data points

predicted values of 0.11 and 0.2. This branch highlights TP and pH as key predictors. The right subtree, where BOD exceeds 0.49, splits based on pH ≤ 8.385 and pH ≤ 8.01, followed by a rainfall split at ≤ 154.85, leading to leaf nodes with predicted values between 0.029 and 0.122. Further splits based on water temperature (≤ 28.6) and silicate (≤ 7.705) refine the predictions, emphasizing the importance of pH, rainfall, water temperature, and silicate concentrations model had an R^2 of 0.60, whereas. MPR model had an R^2 of 0.75 Furthermore, compared to the similar figures from the MPR (RMSE: 0.03, MAE: 0.023) and GAMs (RMSE: 0.045, MAE: 0.033), the DTR model yielded an RMSE of 0.04 and an MAE of 0.029. Additionally, the DTR model's MAPE was greater (76.228) than the MPR's (58.313), suggesting that its forecasts were generally less accurate. Additionally, compared to the MPR model, the DTR model's NSE value was 0.597, indicating that it is less successful at predicting the nitrate concentration. The MPR model's NSE value was 0.754.

Random Forest (RF)

The Random Forest algorithm is the current emphasis. Since Random Forest is a supervised learning method, it can only work with data that has labels or results. It generates several decision trees, each based on a portion of the data that is selected at random. The model then creates an overall forecast for unknown data points by combining the results of each of these decision trees (Sun et al. 2024).

A maximum depth of 20 was used by performing a grid search CV to optimize the Random Forest model, along with 200 estimators, the square root of features for maximum feature selection, a minimum of 1 for leaf samples and a minimum of 5 for split samples. The model performs exceptionally well on the dataset, as evidenced by its MSE of 0.0007 and RMSE of 0.028, which show that the predictions and actual values closely match. Furthermore, a moderate degree of accuracy in percentage terms is demonstrated by the MAPE of 57.27%. Strong predictive ability is indicated

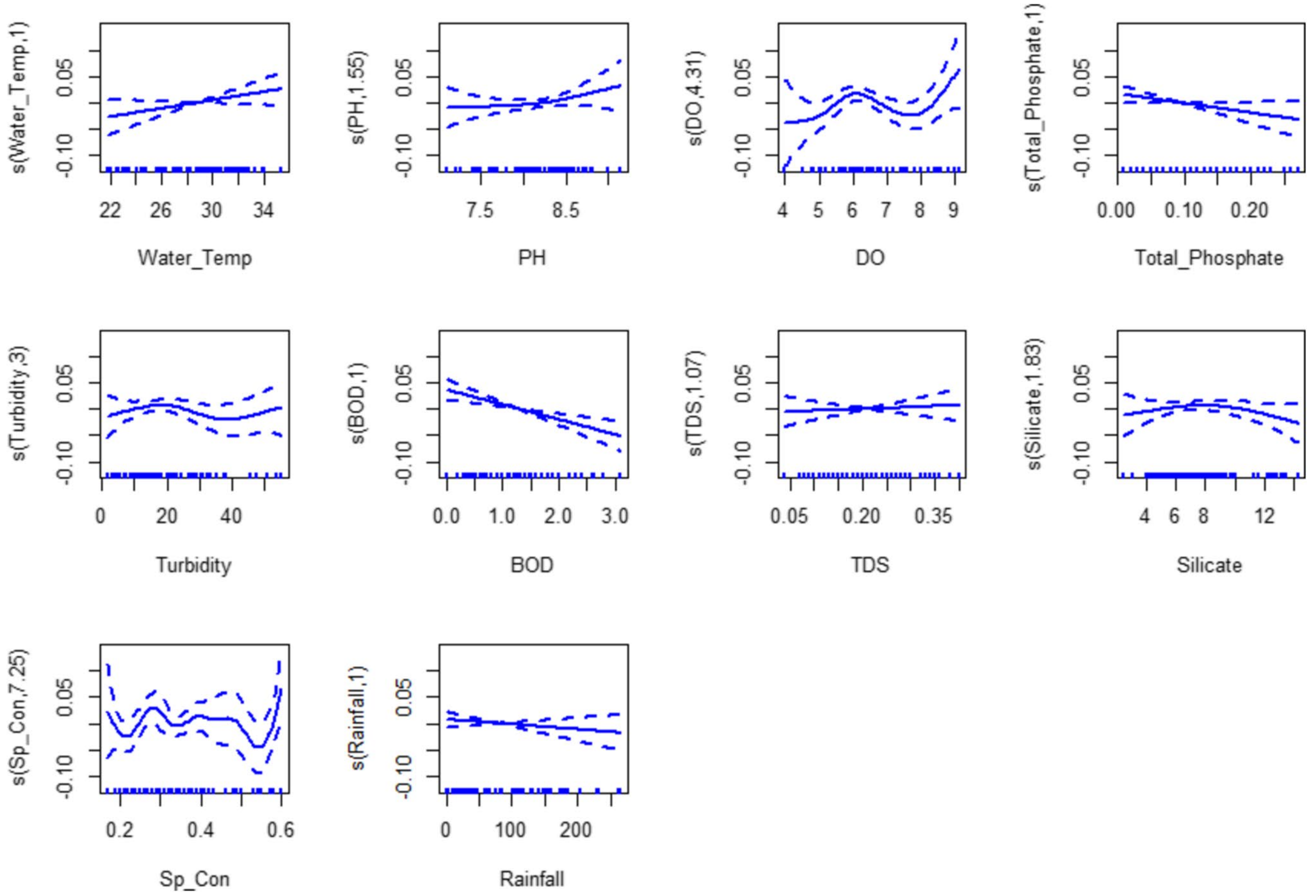

Fig. 6 Generalized Additive Models for nitrate–N Concentration in relation to water quality parameters (Note: Water_Temp: water temp., PH: pH, Total_Phosphate:TP, Sp_Con: specific conductivity)

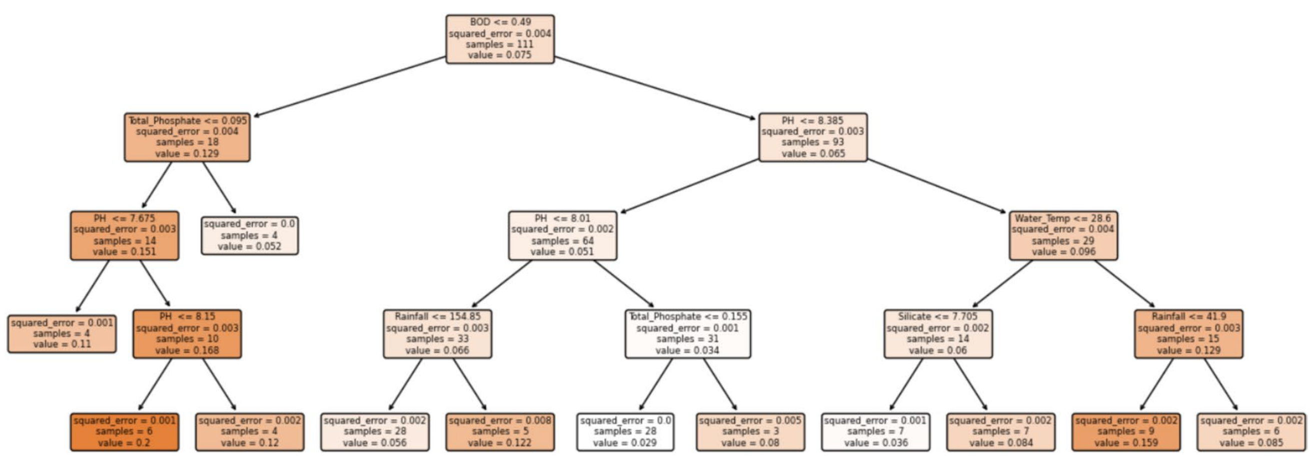

Fig. 7 Decision tree regression for nitrate–N prediction, illustrating hierarchical splits for Biochemical Oxygen Demand (BOD) based on environmental variables, including Total Phosphate, pH, Water_Temp, silicate, and rainfall

by the NSE score of 0.804; values nearer 1 denote higher predictive accuracy. Furthermore, the model's great capacity to uncover underlying patterns in the data is demonstrated by its coefficient of determination (R^2) score of 0.80, which shows that the model explains almost 80% of the variability in the target variable. This shows that the Random Forest

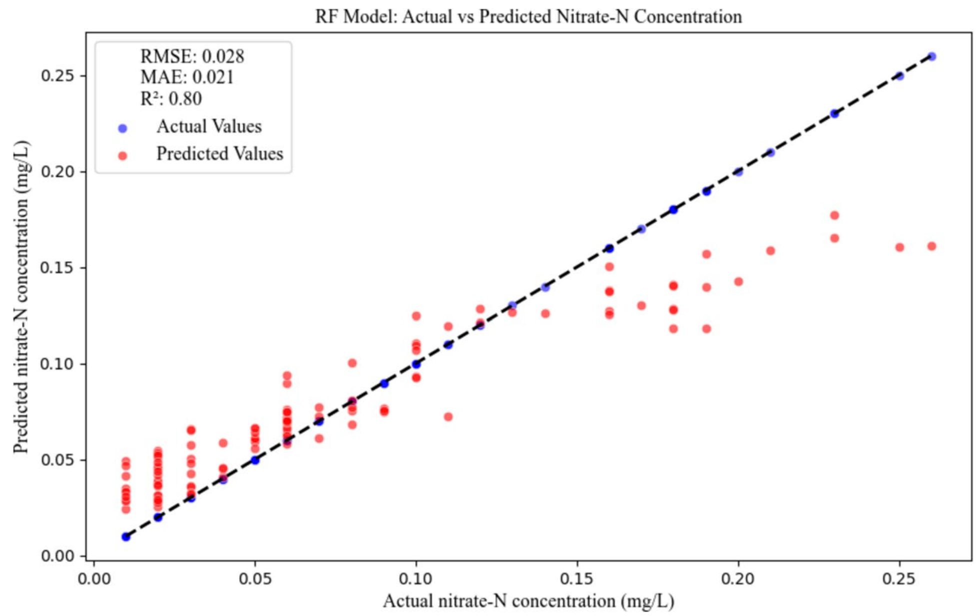

Fig. 8 Comparison of predicted and actual nitrate–N concentrations using a Random Forest model

model is quite successful for the task at hand, providing dependable forecasts with a high degree of accuracy because of its finely adjusted parameters (Fig. 8).

XGBoost (Extreme Gradient Boosting)

An XGBoost Regressor model with particular settings was used further in this study to predict nitrate–N concentration. The model was executed with estimators equal to 5 and five decision trees consecutively trained to ensure an iterative improvement in prediction accuracy. Early stopping was used with early stopping rounds equal to 30 to cease training when the RMSE metric did not show any progress for 30 rounds to prevent overfitting. In addition to improving computing efficiency, this method encouraged the generalization of models. The natural capacity of XGBoost to manage intricate relationships among structured data and its regularization strategies highlighted its appropriateness for the large-scale datasets that worked With an MAE of 0.018 and an RMSE of 0.024, MAPE of 51.805 and NSE of 0.85 the XGBoost model performs effectively in the dataset. These measures show a slight variance between the model's predictions and the actual values (Fig. 9). The model obtains a remarkable R-squared value of 0.85, signifying that it accounts for 85% of the variation in the dependent variable.

Feature importance analysis of predictor variables

This study rigorously examined the key variables influencing nitrate–N concentrations in the lower stretch of the Ganga River, employing three robust machine learning models, i.e., Decision Tree Regression, Random Forest, and XGBoost. Through comprehensive analysis, several variables emerged as pivotal in shaping nitrate–N dynamics in aquatic ecosystems. Figure 10 illustrates the relative importance of predictors for predicting nitrate–N concentrations in the studied stations. While comparing the feature importance in the preformed models, the Biochemical Oxygen Demand stood out as the most significant variable across all models, followed by rainfall, pH and total phosphate. In the present study, temperature was equally crucial for shaping the nitrogen concentrations in the river water.

Model comparison

The performance of the five models—Multiple Polynomial Regression (MPR), Generalized Additive Models (GAMs), Decision Tree Regression, Random Forest, and XGBoost—on the dataset can be compared using their RMSE, MAE, MAPE, NSE, and R^2 values, as summarized in Table 3. XGBoost exhibits the best performance with the lowest RMSE (0.024), MAE (0.018), and MAPE (51.81%), along with the highest NSE (0.855) and R^2 (0.85), indicating it explains 85% of the variance in nitrate–N concentration. The Random Forest model also demonstrates strong performance, achieving an RMSE of 0.028, MAE of 0.021, MAPE of 57.27%, NSE of 0.804, and R^2 of 0.80, showing that it captures 80% of the variance in the data. Multiple Polynomial Regression (MPR) follows with decent accuracy, achieving an RMSE of 0.03, MAE of 0.023, MAPE of 58.31%, NSE of 0.754, and R^2 of 0.75. Generalized Additive Models (GAMs) exhibit moderate performance with an RMSE of 0.045, a MAE of 0.033, a higher MAPE of 82.84%, NSE of 0.485, and R^2 of 0.48, indicating less predictive power compared to other models. The Decision Tree Regression has higher errors, with an RMSE of 0.04

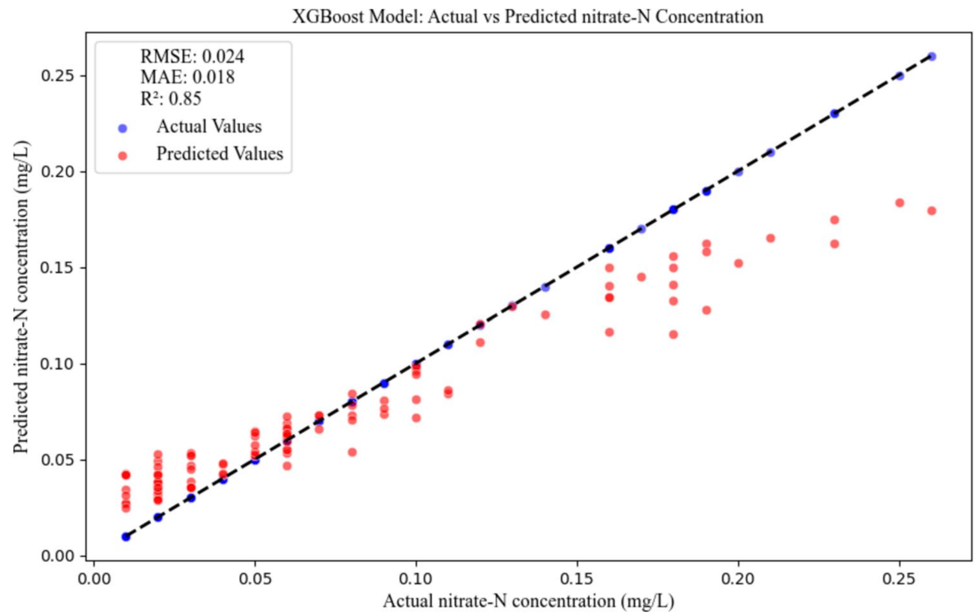

Fig. 9 Predicted versus actual nitrate–N concentrations using an XGBoost model

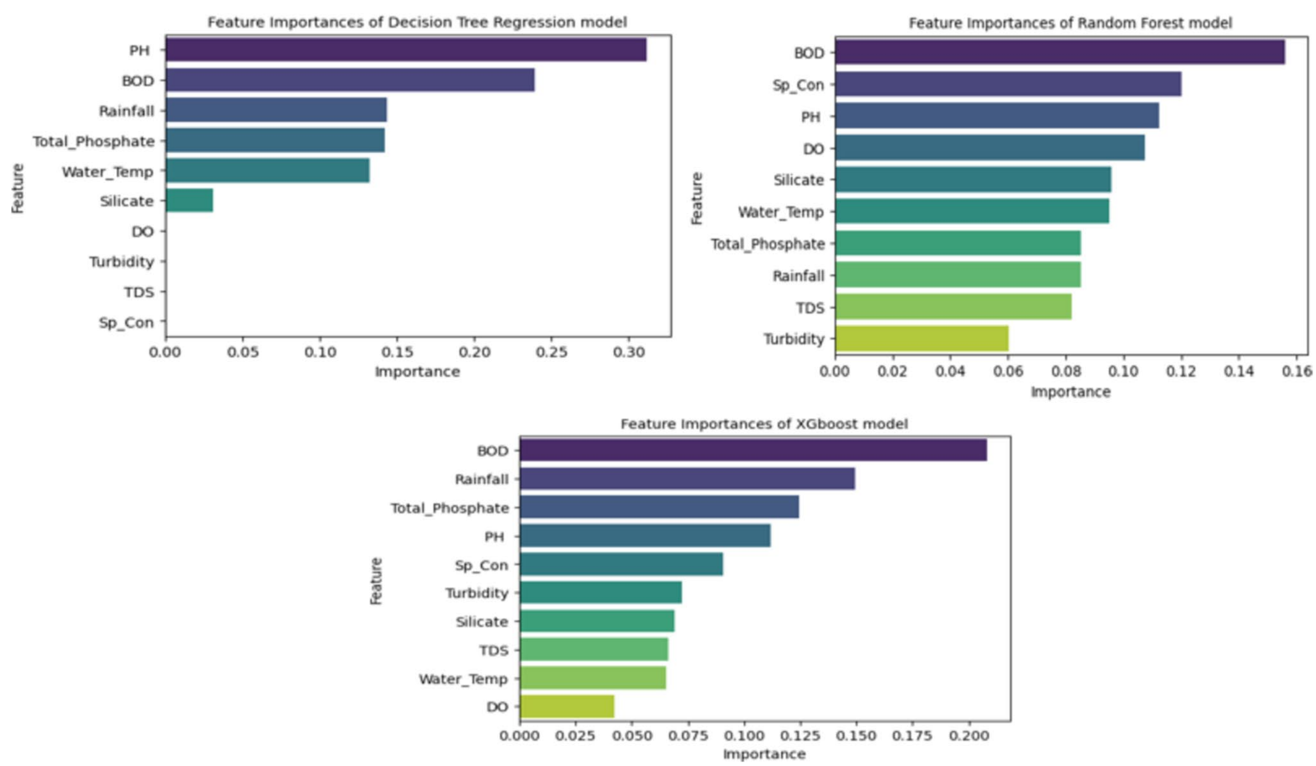

Fig. 10 Relative importance of predictors for DTR, RF, and XGBoost models in predicting nitrate–N concentrations in the lower stretch of the river Ganga

and MAE of 0.029, a MAPE of 76.23%, and a lower NSE of 0.597, resulting in an R^2 of 0.60. This suggests that it explains only 60% of the variance in the data. Overall, XGBoost and Random Forest models emerge as the top-performing models among the five, demonstrating their ability to capture complex relationships within the dataset.

Discussion

In recent decades, the Ganga River has been subjected to anthropogenic pressure due to discharge of untreated sewage and chemical waste (Sengupta et al. 2014; Hamner

Table 3 Performance metrics of the model in the present study

Model	RMSE	MAE	MAPE	NSE	R^2
Multiple Polynomial Regression	0.03	0.023	58.313	0.754	0.75
Generalized Additive Models	0.045	0.033	82.839	0.485	0.48
Decision Tree Regression	0.04	0.029	76.228	0.597	0.60
Random Forest	0.028	0.021	57.272	0.804	0.80
XGboost	0.024	0.018	51.805	0.855	0.85

et al. 2013). This pollution has dire consequences for public health and spreads water borne diseases. (Rai 2013). Abrupt changes in physicochemical parameters including increase pollutants in aquatic ecosystems further reveal a stressed ecosystem, emphasizing the critical need for effective management of both nutrients and pollutants. Industrial discharges from tanneries, sugar and distillery operations, and pulp and paper mills contribute significant loads of Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), total nitrogen (TN), and various metals (Central Pollution Control Board (CPCB), India; Ganga River Basin Environmental Management Plan (GRBEMP), India). Additionally, agricultural runoff exacerbates nutrient pollution, leading to eutrophication fueled by excess nitrogen and phosphorus (Ganga River Basin Environmental Management Plan (GRBMP), India). Despite regulations on effluent disposal, water quality often deteriorates during low-flow periods, highlighting the necessity for stringent monitoring and adaptive management. Elevated organic and microbial loads contribute to reduced dissolved oxygen levels, resulting in fish kills and further ecosystem stress (Hansson and Brönmark 2009; Dodds and Whiles 2020). Addressing these multifaceted challenges through targeted interventions is essential for restoring and sustaining the health of the Ganga River's lower stretch.

The performance of five models—MPR, GAMs, Decision Tree Regression, Random Forest and XGBoost—was evaluated on the dataset using three key metrics—RMSE, MAE, and the coefficient of determination (R^2). These models were chosen because they capture linear and non-linear relationships and comprehensively analyse their predictive capabilities. MPR demonstrated good predictive accuracy with an RMSE of 0.03, a MAE of 0.023, a MAPE of 58.313, an NSE of 0.754 and an R^2 value of 0.75. These metrics indicate that MPR explains 75% of the variance in the nitrate–N concentration, making it a robust choice for modelling non-linear interactions in the data, and it effectively captures the non-linear relationships between nitrate–N concentrations and water quality parameters, providing a significant insight into the underlying dynamics of the dataset. According to (Malone et al. 2007), the MPR model achieved an R^2 of 0.76 for predicting nitrate–N concentration with an RMSE value of 1.90. Knoll et al. (2019) also observed R^2 of 0.42 with RMSE 10.43 and MAE 6.92 for predicting nitrate–N concentration employing Multiple Linear Regression (MLR). In the other study by Mehdaoui et al. (2024) conducted in the Cuyahoga Watershed, the R^2 value was 0.404, the RMSE was 1.0083, and the MAE was 0.7806 for predicting nitrate concentration using MLR.

In the present study, GAMs, which use smooth functions to model non-linear relationships, showed moderate performance. The RMSE was 0.045, the MAE was 0.033, and the R^2 value was 0.48. This indicates that GAM explains 48% of the variance in the target variable. While GAMs provide flexibility in capturing non-linear patterns, their prediction accuracy in this context was lower than that of other models. This performance contrasts Leigh et al. (2019), who achieved less than 22% of deviance explained using Generalized Linear Mixed Models (GLMMs). However, the correlations between nitrate and other water-quality variables tend to be nonlinear, as shown by the present study and Harrison et al. (2021) have noted. In this study, the fitted GAMs model explained a high fraction of variation in nitrate–N (48%) compared with the Generalized Linear Mixed Models (< 22%), which correlates with the work reported by Leigh et al. (2019). Additionally, Kermorvant et al. (2023), Generalized Additive Mixed Models (GAMMs) are well-suited for capturing and interpreting the intricate nonlinear relationships between nitrate levels and various water-quality parameters. These models explain 99% of the variability in nitrate concentration, demonstrating their substantial explanatory capability. From the GAMs model, it was highlighted that DO (p-value = 0.043, ≤ 0.05) and BOD (p-value = 0.000364, ≤ 0.05) were the most significant predictors. Most other predictors did not have significant smooth terms, suggesting that their effects were not statistically significant at this level. The R^2 value for the GAMs model was found to be 48%, which was acceptable but lower than the MPR model. Additionally, an RMSE of 0.045 and a MAE of 0.033 were produced by the GAMs, both slightly lower than those from the MPR. Despite these differences, the GAMs still showed low accuracy but minimal prediction error. These observations align with the findings Kermorvant et al. reported (2023).

In this study, Decision Tree Regression (DTR) showed relatively higher errors with an RMSE of 0.04 and an MAE of 0.029, a MAPE of 76.228, an NSE of 0.597 and the R^2 value of 0.60, indicating that it explains 60% of the variance in the nitrate–N concentration. According to Xue et al. (2024), the Decision Tree Regression model achieved an R^2 of 0.81 for predicting Total Nitrogen (TN) and an R^2 of 0.76 for predicting TP. The recent study of Pokharel and Ghimire (2023) found the same results that the R^2 for Decision Tree models predicting energy consumption in a low-energy house was 0.34, with RMSE and MAE values of 82.28 and 41.18, respectively. Additionally, Mehdaoui et al. (2024) conducted in the Cuyahoga Watershed, the R^2 value was 0.75, the RMSE was

0.5894, and the MAE was 0.4259 for predicting nitrate concentration. For predicting phosphorus using DTR, the R^2 value was 0.55, the RMSE was 0.1258, and the MAE was 0.0801.

The Random Forest model, an ensemble learning technique, performed well in this study with an RMSE of 0.028, an MAE of 0.021, a MAPE of 57.272, an NSE of 0.804 and an R^2 value of 0.80. This indicates that Random Forest explains 80% of the variance in the nitrate–N concentration. As per Harrison et al. (2021), Random Forest Regression models can account for 89% of the variation in nitrate concentration and can effectively capture nonlinear interactions among variables related to the present study. Wheeler et al. (2015) reported the same that the R^2 and RMSE of an RF model pertaining to domestic wells in Iowa were 0.38 and 1.55, respectively. In another study, Knoll et al. (2019), Mehdaoui et al. (2024), and Xue et al. (2024) found similar results employing the Random Forest prediction model.

XGBoost, a scalable and optimized version of gradient boosting, exhibited the best performance among the evaluated models in this study. It achieved the lowest RMSE of 0.024, the lowest MAE of 0.018, the lowest MAPE of 51.805, highest NSE value of 0.855 and the highest R^2 value of 0.85. This indicates that XGBoost explains 85% of the variance in the nitrate–N concentration. Ransom et al. (2022) reported that the XGBoost model for predicting nitrate concentration achieved an R^2 of 0.83 and an RMSE of 1.15 in their study as the best performance model, further validating the effectiveness of XGBoost in various scenarios. Pokharel and Ghimire (2023) reported that the R^2 for Random Forest models predicting energy consumption in a low-energy house was 0.47, with RMSE and MAE values of 79.11 and 37.28, respectively. They also found that the R^2 for XGBoost models was 0.61, with RMSE and MAE values of 65.28 and 29.81, respectively. Similar studies that used the RF and XGBoost algorithms reported that XGBoost performed the best, although the difference between it and RF was minimal (Kiangala and Wang 2021; Peiró-Signes et al. 2022). Since XGBoost and RF are both ensemble algorithms, their predictions are difficult to interpret, and each has unique drawbacks. XGBoost typically produces less accurate results when handling imbalanced data (Kiangala and Wang 2021), while RF's performance is highly influenced by the amount of data used in the training dataset (Pokharel and Ghimire 2023). Wheeler et al. (2015) and Knoll et al. (2019) provide context to the robustness of the Random Forest model across different datasets and geographical scales. The other model performance matrices between present study and previous studies are shown in Table 4.

Variable importance Influencing nitrate–N concentrations in the river water

Based on the results of the above-stated models, the BOD ranked as the most significant variable across all models, followed by rainfall, pH and total phosphate. In the present study temperature, DO, and turbidity were also found to be equally important for shaping the nitrate–N concentrations in the river water. BOD directly impacts nitrate–N levels by influencing dissolved oxygen, which is crucial for nitrification and denitrification processes (Zainol et al. 2022; Huang et al. 2024). The elevated BOD in the river water potentially indicates heightened organic material breakdown.

Rainfall facilitates the transport of pollutants, including organic nitrogen from pit latrines, drains, and agricultural runoff, into surface water bodies. While increased rainfall generally leads to higher nitrate levels in shallow groundwater, its effects on surface water can vary. For instance, higher average precipitation may promote nitrogen absorption by crops or dilute nitrate-containing materials, thereby potentially reducing surface water nitrate levels (Bowes et al. 2015); pH emerged as the third-most influential variable. The pH of surface water significantly influences nitrate–N concentrations by affecting microbial activity crucial for nitrification and denitrification (Zoppas et al. 2024). Optimal pH conditions (typically 6.5 to 8.5) promote nitrification, enhancing nitrate–N removal from water. Deviations from this range can hinder nitrifying microorganisms, leading to decreased nitrate–N uptake by aquatic organisms and potentially increasing nitrate–N levels (Lopez et al. 2024).

Total Phosphate rank was prominent across all performed models. Phosphates from agricultural runoff and wastewater stimulate algal growth in water bodies. Initially, algae absorb nitrogen, reducing nitrate–N levels. However, algae release nitrogen as nitrates back into the water upon decomposition, contributing to increased nitrate–N concentrations. Effective management of phosphate levels is critical for regulating nutrient dynamics and preserving water quality in aquatic ecosystems (Szejba et al. 2016). Likewise, water temperature significantly influences nitrate–N levels through its impact on microbial activities in the nitrogen cycle. Higher temperatures accelerate microbial processes, such as nitrification and denitrification, potentially increasing or decreasing nitrate–N concentrations depending on oxygen availability. Elevated temperatures also promote algal growth, influencing nitrate–N uptake for photosynthesis. Conversely, lower temperatures slow down these processes, leading to higher nitrate–N retention in water bodies (Mustafa et al. 2024).

Conclusion and way forward

The study demonstrates that building a model for nitrate–N using the same set of explanatory water-quality variables is achievable, even for sites with diverse environmental and climatic characteristics, supporting comprehensive knowledge of nitrate–N dynamics both geographically and temporally, as well as well-informed management

Table 4 Comparison of present study with previous studies according to various performance metrics

Previous study	Models				
	Multiple Polynomial Regression (MPR)	Generalized Additive Models (GAMs)	Decision Tree Regression (DTR)	Random Forest (RF)	XGBoost
Malone et al. (2007)	R2 = 0.76, RMSE = 1.90	-	-	-	-
Knoll et al. (2019)	R2 = 0.42, RMSE = 10.43, MAE = 6.92	-	-	-	-
Mehdaoui et al. (2024)	R2 = 0.404, RMS:1.008; MAE = 0.780	-	R^2 = 0.75, RMSE = 0.5894, MAE = 0.4259	-	-
Leigh et al. (2019)	-	R2 = < 22%		-	-
Kermorvant et al. (2023)	-	R2 = 99%		-	-
Xue et al. (2024)	-	-	R2 (TN) = 0.81, R2 (TP) = 0.76	-	-
Pokharel and Ghimire (2023)	-	-	R2 = 0.34, RMSE = 82.28, MAE = 41.18	-	R^2 = 0.61, RMSE = 65.28, MAE = 29.81
Harrison et al. (2021)	-	-	-	R^2 = 0.89	
Wheeler et al. (2015)	-	-	-	R^2 = 0.38, RMSE = 1.55	
Ransom et al. (2022)	-	-	-	-	R^2 = 0.83, RMSE = 1.15
Present study (2024)	RMSE = 0.03, MAE = 0.023, MAPE = 58.313, NSE = 0.754, R^2 = 0.75	RMSE = 0.045, MAE = 0.033, MAPE = 82.839, NSE = 0.485, R^2 = 0.48	RMSE = 0.04, MAE = 0.029, MAPE = 76.228, NSE = 0.597, R^2 = 0.60	RMSE = 0.028, MAE = 0.021, MAPE = 57.272, NSE = 0.804, R^2 = 0.80	RMSE = 0.024, MAE = 0.018, MAPE = 51.805, NSE = 0.855, R^2 = 0.85

strategies. Also, the present study evaluated the predictive performance of five machine learning models: Polynomial Regression (MPR), Generalized Additive Models (GAMs), Decision Tree Regression, Random Forest, and XGBoost, on nitrate–N concentration data from the lower stretch of the Ganga River, India. Further, it demonstrated the efficacy of advanced machine learning models in predicting nitrate–N concentrations in the lower stretch of the Ganga River. Among the evaluated models, XGBoost and Random Forest emerged as the most effective, explaining 85% and 80% of the variance in the target variable, respectively. Multiple Polynomial Regression (MPR) also showed good performance, while Generalized Additive Models and Decision Tree Regression had lower predictive capabilities. These results highlight the varying effectiveness of different models in capturing non-linear relationships and complex interactions among water quality parameters. Comparative analysis with previous studies underscores the robustness of XGBoost and Random Forest in diverse environmental contexts. The insights gained from this study highlight the potential of machine learning models in ecological monitoring and management, paving the way for more informed and proactive water quality management strategies. From the studied models, it is evident that the discrepancies of nitrate–N concentration were primarily influenced by the water variables such as BOD, pH, Rainfall, water temperature, and total phosphate. In addition, the nitrate–N concentration was within the permissible limit of Bureau of Indian Standards (BIS 2011; Aswal 2020), indicating a healthy scenario in the studied stretch without nitrate–N pollution thereof. Future research should focus on incorporating additional environmental variables, such as phosphorus, sulphur, and other heavy metals, and exploring ensemble methods to further enhance model performance and robustness. By continuously improving predictive models, it becomes possible to understand better and mitigate the impacts of nitrate–N pollution, ultimately contributing to the sustainability of vital water resources like the mighty Ganga River.

Acknowledgements Authors are grateful to the ministry of Jal Sakti, Government of India, under the Namami Ganga Mission-II, Fish conservation and stock enhancement of fishery of Ganga River basin project (Ad-35012/1/2023-NMCG-NMCG) for financial support to carry out the research work.

Authors' contribution B. K. Das: Supervision, conceptualization and resource. S. Paul: Conceptualization, data curation, data analysis, writing original draft; B. Mandal: Conceptualization, data curation, interpretation of data and writing original draft; P. Gogoi:

Conceptualization, interpretation of data, review & editing, and manuscript preparation; Liton Paul: Map preparation and writing, and editing. A. Saha: Review & editing and data interpretation. C. Johnson: data curation and data interpretation. Akankshya Das: data curation and manuscript preparation. Archisman Ray: formal analysis, data interpretation and manuscript preparation. Shreya Roy: data curation and manuscript preparation. Shubhadeep Das Gupta: data curation and manuscript preparation.

Funding The funding for the study was provide by the ministry of Jal Shakti, Government of India, under the Namami Ganga Mission-II, Fish conservation and stock enhancement of fishery of Ganga River basin project (Ad-35012/1/2023-NMCG-NMCG).

Data availability Data will be made available on request to corresponding author/ first author.

Declarations

Ethical statement The research was done according to ethical standards.

Consent to participate Not applicable.

Consent to publish In this study, does not have any individual data.

Competing interests All the authors declare no competing interests.

Declaration of generative AI While preparation of manuscript, no AI tools have been used to analyses and draw insights from data as part of the research process.

References

Abba SI, Yassin MA, Jibril MM, Tawabini B, Soupios P, Khogal A, Shah SMH, Usman J, Aljundi IH (2024) Nitrate concentrations tracking from multi-aquifer groundwater vulnerability zones: insight from machine learning and spatial mapping. Process Saf Environ Prot 184:1143–1157. https://doi.org/10.1016/j.psep.2024.02.041

Agbasi JC, Egbueri JC (2024) Prediction of potentially toxic elements in water resources using MLP-NN, RBF-NN, and ANFIS: a comprehensive review. Environ Sci Pollut Res 1–29. https://doi.org/10.1007/s11356-024-33350-6

APHA (2012) Standard methods for the examination of water and wastewater. In Rice EW, Baird RB, Eaton AD, Clesceri LS (Eds.) American Public Health Association (APHA), American Water Works Association (AWWA) and Water Environment Federation (WEF) (22nd ed.). Washington DC

Aslan V (2024) The analysis of classical, polynomial regression and cubic spline mathematical models in hemp biodiesel optimization: an experimental comparison. Environ Sci Pollut Res 31:9392–9940. https://doi.org/10.1007/s11356-023-31720-0

Aswal DK (2020) Quality infrastructure of India and its importance for inclusive national growth. Mapan 35(2):139–150. https://doi.org/10.1007/s12647-020-00376-3

Attri SD, Tyagi A (2010) Climate profile of India. Environment Monitoring and Research Center, India Meteorology Department: New Delhi, India. https://doi.org/10.22541/essoar.171405637.76928549/v1

Banik S, Balasubramanian K, Manna S, Derrible S, Sankaranarayananan SK (2024) Evaluating generalized feature importance via performance assessment of machine learning models for predicting elastic properties of materials. Comput Mater Sci 236:112847. https://doi.org/10.1016/j.commatsci.2024.112847

Belany P, Hrabovsky P, Sedivy S, CajovaKantova N, Florkova Z (2024) A comparative analysis of polynomial regression and artificial neural networks for predictionof lighting consumption. Buildings 14(6):1712. https://doi.org/10.3390/buildings14061712

Bhatt S, Mishra AP, Chandra N, Sahu H, Chaurasia SK, Pande CB, Agbasi JC, Khan MYA, Abba SI, Egbueri JC, Đurin B (2024) Characterizing seasonal, environmental and human-induced factors influencing the dynamics of Rispana River's water quality: Implications for sustainable river management. Results Eng 22:102007. https://doi.org/10.1016/j.rineng.2024.102007

Bhattara A, Dhakal S, Gautam Y, Bhattarai R (2021) Prediction of nitrate and phosphate concentrations using machine learning algorithms in watersheds with different landuse. Water 13(21):3096. https://doi.org/10.3390/w13213096

BIS (2012) Bureau of Indian Standards: drinking water specification (Second revision), IS 10500:2012

Bouke MA, Zaid SA, Abdullah A (2024) Implications of data leakage in machine learning preprocessing: a multi-domain investigation. https://doi.org/10.21203/rs.3.rs-4579465/v1

Boumans L, Fraters D, van Drecht G (2004) Nitrate leaching by atmospheric N deposition to upper groundwater in the sandy regions of the Netherlands in 1990. Environ Monit Assess 93:1–15. https://doi.org/10.1023/B:EMAS.0000016788.24386.be

Bowes MJ, Jarvie HP, Halliday SJ, Skeffington RA, Wade AJ, Loewenthal M, Gozzard E, Newman JR, Palmer-Felgate EJ (2015) Characterising phosphate and nitrate inputs to a rural river using high-frequency concentration–flow relationships. Sci Total Environ 511:608–620. https://doi.org/10.1016/j.scitotenv.2014.12.086

Bricker SB, Longstaff B, Dennison W, Jones A, Boicourt K, Wicks C, Woerner J (2008) Effects of nutrient enrichment in the nation's estuaries: a decade of change. Harmful Algae 8(1):21–32. https://doi.org/10.1016/j.hal.2008.08.028

Camargo JA, Alonso A, Salamanca A (2005) Nitrate-N toxicity to aquatic animals: a review with new data for freshwater invertebrates. Chemosphere 58(9):1255–1267. https://doi.org/10.1016/j.chemosphere.2004.10.044

Canter LW (2019) Nitrates in groundwater. Routledge. https://doi.org/10.1201/9780203745793

Chakraborty M, Sriram V, Murali K (2024) Investigation of ship-induced hydrodynamics and sediment resuspension in a restricted waterway. Appl Ocean Res 142:103831. https://doi.org/10.1016/j.apor.2023.103831

Choudhury A, Mondal A, Sarkar S (2024) Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: A comparative study and review of Random Forest, Adaboost, XGboost and LightGBM frameworks. Eur Phys J Spec Top 1–39. https://doi.org/10.48550/arXiv.2405.06040

Dodds WK, Whiles MR (2020) Microbes and plants. Freshw Ecol 211–249

Dovers E, Stoklosa J, Warton DI (2024) Fitting log-Gaussian Cox processes using generalized additive model software. Am Statist 13:1–16. https://doi.org/10.1080/00031305.2024.2316725

Duc L, Sawada Y (2023) A signal-processing-based interpretation of the Nash-Sutcliffe efficiency. Hydrol Earth Syst Sci 27(9):1827–1839

Dupas R, Faucheux M, Kiessé TS, Casanova A, Brekenfeld N, Fovet O (2024) High-intensity rainfall following drought triggers extreme nutrient concentrations in a small agricultural catchment. Water Res 264:122108. https://doi.org/10.1016/j.watres.2024.122108

Egbueri JC, Abu M, Agbasi JC (2024) An integrated appraisal of the hydro-geochemistry and the potential public health risks of groundwater nitrate and fluoride in eastern Ghana. Groundw Sustain Dev 26:101264

Elavarasan D, Vincent PDR (2021) A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. J Ambient Intell Hum Comput 12(11):10009–10022. https://doi.org/10.1007/s12652-020-02752-y
Fenice M (2021) The nitrogen cycle: An overview. In: Nitrogen Cycle, pp 1–21. https://doi.org/10.1201/9780429291180-1
Gao Y, Fang Z, Van Zwieten L, Bolan N, Dong D, Quin BF, Meng J, Li F, Wu F, Wang H, Chen W (2022) A critical review of biochar-based nitrogen fertilizers and their effects on crop production and the environment. Biochar 4(1):36
Ghanizadeh AR, Amlashi AT, Bahrami A, Isleem HF, Dessouky S (2024) A formulation for asphalt concrete air void during service life by adopting a hybrid evolutionary polynomial regression and multi-gene genetic programming. Sci Rep 14(1):13254. https://doi.org/10.1038/s41598-024-61313-x
Green PA, Vörösmarty CJ, Meybeck M, Galloway JN, Peterson BJ, Boyer EW (2004) Pre-industrial and contemporary fluxes of nitrogen through rivers: a global assessment based on typology. Biogeochemistry 68:71–105
Hamner S, Pyke D, Walker M, Pandey G, Mishra RK, Mishra VB (2013) Sewage pollution of the River Ganga: an ongoing case study in Varanasi. India River Syst 20(3–4):157–167. https://doi.org/10.1127/1868-5749/2013/0058
Hansson A, Brönmark C (2009) Biomanipulation of aquatic ecosystems. Lake Ecosystem Ecology: A Global Perspective 396
Harrison JW, Lucius MA, Farrell JL, Eichler LW, Relyea RA (2021) Prediction of stream nitrogen and phosphate concentrations from high-frequency sensors using Random Forests Regression. Sci Total Environ 763:143005. https://doi.org/10.1016/j.scitotenv.2020.143005
Heddam S, Kim S, Elbeltagi A, Malik A, Zounemat-Kermani M, Kisi O (2023) Predicting nitrate concentration in river using advanced artificial intelligence techniques: extreme learning machines versus deep learning. In:Water, Land, and Forest Susceptibility and Sustainability, Elsevier, pp 121–153. https://doi.org/10.1016/B978-0-323-91880-0.00030-1
Howarth R, Swaney D, Billen G, Garnier J, Hong B, Humborg C, Johnes P, Mörth CM, Marino R (2012) Nitrogen fluxes from the landscape are controlled by net anthropogenic nitrogen inputs and by climate. Front Ecol Environ 10(1):37–43
Huang Y, Li L, Li R, Li B, Wang Q, Song K (2024) Nitrogen cycling and resource recovery from aquaculture wastewater treatment systems: a review. Environ Chem Lett 22:2467–2482. https://doi.org/10.1007/s10311-024-01763-x
Jaddi H, El-Hmaidi A, Ousmana H, Berrada M, Aouragh MH, Iallamen Z, Kasse Z, El Ouali A, Boufal MH, Ragragui H, Saouita J (2024) Predicting nitrate levels in the saïss water table: a comparative study of machine learning methods, BIO web of conferences. EDP Sciences 115:03001. https://doi.org/10.1051/bioconf/202411503001
Kermorvant C, Liquet B, Litt G, Mengersen K, Peterson EE, Hyndman RJ, Jones JB Jr, Leigh C (2023) Understanding links between water-quality variables and nitrate concentration in freshwater streams using high frequency sensor data. PLoS One 18(6):e0287640. https://doi.org/10.1371/journal.pone.0287640
Khalifa FA, Abdelkader HM, Elsaid AH (2024) An analysis of ensemble pruning methods under the explanation of Random Forest. Inf Syst 120:102310
Kiangala SK, Wang Z (2021) An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. Mach Learn Appl 4:100024. https://doi.org/10.1016/j.mlwa.2021.1000245
Knoll L, Breuer L, Bach M (2019) Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci Total Environ 668:1317–1327. https://doi.org/10.1016/j.scitotenv.2019.03.045
Korostynska O, Mason A, Al-Shamma'a A (2012) Monitoring of nitrates and phosphates in wastewater: current technologies and further challenges. Int J Smart Sens Intell Syst 5(1):149–176. https://doi.org/10.21307/ijssis-2017-475
Kumar L, Afzal MS, Ahmad A (2022) Prediction of water turbidity in a marine environment using machine learning: a case study of Hong Kong. Reg Stud Mar Sci 52:102260. https://doi.org/10.1016/j.rsma.2022.102260
Kumar S, Saxena A, Srivastava RK, Singh SB, Ram RN, Pandey NN (2024) Heavy metals in fishes, water and macrophyte of the Ganga River and risk related to their consumption. Toxicol Int 379–390. https://doi.org/10.18311/ti/2024/v31i3/36636
Kuypers MM, Marchant HK, Kartal B (2018) The microbial nitrogen-cycling network. Nat Rev Microbiol 16(5):263–276. https://doi.org/10.1038/nrmicro.2018.9
Lai J, Tang J, Li T, Zhang A, Mao L (2024) Evaluating the relative importance of predictors in Generalized Additive Models using the gam. hp R Package Plant Divers 46:542–546. https://doi.org/10.1016/j.pld.2024.06.002
Lamontagne JR, Barber CA, Vogel RM (2020) Improved estimators of model performance efficiency for skewed hydrologic data. Water Resour Res 56(9):e2020WR027101
Lee J, Sim MK, Hong JS (2024) Assessing decision tree stability: a comprehensive method for generating a stable decision tree. IEEE Access 12:90061–90072. https://doi.org/10.1109/ACCESS.2024.3419228
Leigh C, Kandanaarachchi S, McGree JM, Hyndman RJ, Alsibai O, Mengersen K, Peterson EE (2019) Predicting sediment and nutrient concentrations from high-frequency water-quality data. PLoS One 14(8):e0215503. https://doi.org/10.1371/journal.pone.0215503
Li X, Xu YJ, Ni M, Wang C, Li S (2023) Riverine nitrate source and transformation as affected by land use and land cover. Environ Res 222:115380
Lopez K, Leme VF, Warzecha M, Davidson PC (2024) Wastewater nutrient recovery via fungal and nitrifying bacteria treatment. Agriculture 14(4):580. https://doi.org/10.3390/agriculture14040580
Mahboobi H, Shakiba A, Mirbagheri B (2023) Improving groundwater nitrate concentration prediction using local ensemble of machine learning models. J Environ Manag 345:118782. https://doi.org/10.1016/j.jenvman.2023.118782
Mahlknecht J, Torres-Martínez JA, Kumar M, Mora A, Kaown D, Loge FJ (2023) Nitrate prediction in groundwater of data scarce regions: the futuristic fresh-water management outlook. Sci Total Environ 905:166863. https://doi.org/10.1016/j.scitotenv.2023.166863
Malhotra M, Dar AA, Jain A, Adithya CV (2024) Unveiling the power of machine learning algorithms. In machine learning and data science techniques for effective government service delivery. IGI Global 114–156. https://doi.org/10.4018/978-1-6684-9716-6.ch005
Malone RW, Ma L, Karlen DL, Meade T, Meek D, Heilman P, Kanwar RS, Hatfield JL (2007) Empirical analysis and prediction of nitrate loading and crop yield for corn–soybean rotations. Geoderma 140(3):223–234. https://doi.org/10.1016/j.geoderma.2007.04.007
Manzali Y, Akhiat Y, Abdoulaye Barry K, Akachar E, El Far M (2024) Prediction of student performance using random forest combined with Naïve Bayes. Comput J p.bxae 036
McDowell RW, McNeill SJ, Drewry JJ, Law R, Stevenson B (2024) Difficulties in using land use pressure and soil quality indicators to predict water quality. Sci Total Environ 935:173445. https://doi.org/10.1016/j.scitotenv.2024.173445
Mehdaoui I, Boudibi S, Latif SD, Sakaa B, Chaffai H, Hani A (2024) Prediction of nitrate concentrations using multiple linear regression and radial basis function neural network in the Cheliff River

basin. Algeria J Appl Water Eng Res 12(1):77–89. https://doi.org/10.1080/23249676.2023.2207838

Modhugu VR, Ponnusamy S (2024) Comparative analysis of machine learning algorithms for liver disease prediction: SVM, logistic regression, and decision tree. Asian J Res Comput Sci 17(6):188–201. https://doi.org/10.9734/ajrcos/2024/v17i6467

Moore AP, Bringolf RB (2018) Effects of Nitrate-N on freshwater mussel glochidia attachment and metamorphosis success to the juvenile stage. Environ Pollut 242:807–813. https://doi.org/10.1016/j.envpol.2018.07.047

Murphy NP (2022) Examining Nitrate Leaching Potential and Nitrogen Cycle Dynamics under Agricultural Managed Aquifer Recharge in the Central Valley of California. University of California, Davis

Mustafa G, Chretien J, Chappell P, Bate A, Ha K, Smith A (2024) Effect of nitrite and nitrate-N accumulation and removal on the suspended solids in the aeration tank of Opal's secondary water treatment plant. Int J Agric Environ Sci 11(1):1–6. https://doi.org/10.14445/23942568/IJAES-V11I1P101

Nair JP, Vijaya MS (2022) River water quality prediction and index classification using machine learning. In J Phys Conf 2325(1):012011

NaseriBoroujeni S, Maribo-Mogensen B, Liang X, Kontogeorgis GM (2024) Novel model for predicting the electrical conductivity of multisalt electrolyte solutions. J Phys Chem B 128(2):536–550

Negru D (2024) Converting Decision Trees into Fault Trees, Bachelor's thesis, University of Twente. https://purl.utwente.nl/essays/100951. Accessed 10.12.2024

Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN (2024) Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics 14(2):144. https://doi.org/10.3390/diagnostics14020144

Özdemir R, Taşyürek M, Aslantaş V (2024) Improved marine predators algorithm and extreme gradient boosting (XGBoost) for shipment status time prediction. Knowl-Based Syst 294:111775. https://doi.org/10.1016/j.knosys.2024.111775

Parhi SK, Patro SK (2024) Compressive strength prediction of PET fiber-reinforced concrete using Dolphin echolocation optimized decision tree-based machine learning algorithms. Asian J Civil Eng 25(1):977–996. https://doi.org/10.1007/s42107-023-00826-8

Patra A, Bhaskaran PK (2016) Trends in wind-wave climate over the head Bay of Bengal region. Int J Climatol 36(13):4222–4240

Patra AK, dos Santos Ribeiro LP, Yirga H, Puchala R, Goetsch AL (2024) Influence of the concentration and nature of total dissolved solids in brackish groundwater on water intake, nutrient utilization, energy metabolism, ruminal fermentation, and blood constituents in different breeds of mature goats and sheep. Sci Total Environ 907:167949. https://doi.org/10.1016/j.scitotenv.2023.167949

Peiró-Signes Á, Segarra-Ona M, Trull-Domínguez Ó, Sánchez-Planelles J (2022) Exposing the ideal combination of endogenous–exogenous drivers for companies' ecoinnovative orientation: Results from machine-learning methods. Socio-Econ Plan Sci 79:101145

Pokharel S, Ghimire P (2023) Data-driven ML models for accurate prediction of energy consumption in a low-energy house: a comparative study of XGBoost, random forest, decision tree, and support vector machine. J Innov Eng Educ 6(1):12–20. https://doi.org/10.3126/jiee.v6i1.54965

Rai B (2013) Pollution and conservation of Ganga River in modern India. Int J Sci Res Publ 3(4):1–4

Ransom KM, Nolan BT, Stackelberg PE, Belitz K, Fram MS (2022) Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. Sci Total Environ 807:151065. https://doi.org/10.1016/j.scitotenv.2021.151065

Salman HA, Kalakech A, Steiti A (2024) Random forest algorithm overview. Baby J Mach Learn 69–79. https://doi.org/10.58496/BJML/2024/007

Schiller E, Müller S, Ebertsch K, Stczhöfer JP (2024) No data left behind: exogenous variables in long-term forecasting of nursing staff capacity. In: 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp 1–10

Senarathne S, van Geldern R, Chandrajith R, Barth JA (2024) Unexpected contributions by carbonates and organic matter in a silicate-dominated tropical catchment: an isotope approach. Sci Total Environ 948:174651. https://doi.org/10.1016/j.scitotenv.2024.174651

Sengupta C, Sukumaran D, Barui D, Saha R, Chattopadhyay A, Naskar A, Dave S (2014) Water health status in lower reaches of river Ganga, India. Appl Ecol Environ Sci 2:20–24. https://doi.org/10.12691/aees-2-1-3

Shaik NB, Jongkittinarukorn K, Bingi K (2024) XGBoost based enhanced predictive model for handling missing input parameters: a case study on gas turbine. Case Stud Chem Environ Eng 10:100775. https://doi.org/10.1016/j.cscee.2024.100775

Shams MY, Elshewey AM, El-Kenawy ESM, Ibrahim A, Talaat FM, Tarek Z (2024) Water quality prediction using machine learning models based on grid search method. Multimed Tools Appl 83(12):35307–35334. https://doi.org/10.1007/s11042-023-16737-4

Siems J, Ditschuneit K, Ripken W, Lindborg A, Schambach M, Otterbach J, Genzel M (2024) Curve your enthusiasm: concurvity regularization in differentiable generalized additive models. Adv Neural Inf Process Syst 36:19029–19057

Singh G, Singh J, Wani OA, Egbueri JC, Agbasi JC (2024) Assessment of groundwater suitability for sustainable irrigation: a comprehensive study using indexical, statistical, and machine learning approaches. Groundw Sustain Dev 24:101059

Singh V (2024) Solid waste management. In:Textbook of Environment and Ecology, Springer Nature Singapore, Singapore, pp 299–307. https://doi.org/10.1007/978-981-99-8846-4_21

Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X (2024) An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst Appl 237:121549. https://doi.org/10.1016/j.eswa.2023.121549

Szejba D, Papierowska E, Cymes I, Bankowska A (2016) Nitrate nitrogen and phosphate concentrations in drainflow: an example of clay soil. J Elementol 21(3). https://doi.org/10.5601/jelem.2015.20.4.922

Visser L, Pat-El R, Lataster J, Van Lankveld J, Jacobs N (2024) Beyond difference scores: unlocking insights with polynomial regression in studies on the effects of implicit-explicit congruency. Psychol Belgica 64(1):5. https://doi.org/10.5334/pb.1246

Wang X, Li Y, Qiao Q, Tavares A, Liang Y (2023) Water quality prediction based on machine learning and comprehensive weighting methods. Entropy 25(8):1186

Wang Z, Huang S, Mu Z, Leng G, Duan W, Ling H, Xu J, Zheng X, Li P, Li Z, Guo W (2024) Relative humidity and solar radiation exacerbate snow drought risk in the headstreams of the Tarim River. Atmos Res 297:107091. https://doi.org/10.1016/j.atmosres.2023.107091

Wheeler DC, Nolan BT, Flory AR, DellaValle CT, Ward MH (2015) Modeling groundwater nitrate concentrations in private wells in Iowa. Sci Total Environ 536:481–488. https://doi.org/10.1016/j.scitotenv.2015.07.080

Xue J, Yuan C, Ji X, Zhang M (2024) Predictive modeling of nitrogen and phosphate concentrations in rivers using a machine learning framework: a case study in an urban-rural transitional area in Wenzhou China. Sci Total Environ 910:168521. https://doi.org/10.1016/j.scitotenv.2023.168521

Yotsuyanagi H, Morohashi M, Takahashi M, Ohizumi T, Inomata Y, Yabusaki S, Tayasu I, Okochi H, Sase H (2024) Sulfate runoff processes during rainfall events in a small, forested catchment on the Sea of Japan side recovering from acidification under climate change. Hydrol Process 38(7):e15221. https://doi.org/10.1002/hyp.15221

Zainol N, Samad KA, Jazlan CAIC, Razahazizi NA (2022) Optimization of COD, nitrate-N and phosphate removal from hatchery wastewater with acclimatized mixed culture. Heliyon 8(4):e09217. https://doi.org/10.1016/j.heliyon.2022.e09217

Zhang D, Wang P, Cui R, Yang H, Li G, Che A, Wang H (2022) Electrical conductivity and dissolved oxygen as predictors of nitrate concentrations in shallow groundwater in Erhai Lake region. Sci Total Environ 802:149879. https://doi.org/10.1016/j.scitotenv.2021.149879

Zoppas FM, Sacco N, Beltrame T, Akhter F, Miró E, Marchesini FA (2024) Maximizing selectivity and activity in the catalytic reduction of nitrates with formic acid under optimal pH conditions. Next Sustain 3:100030. https://doi.org/10.1016/j.nxsust.2024.100030