International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I1P130
Volume 72 Issue 1, 311-324, January 2024

Original Article

A Novel Method to Predict the Nitrate Concentration Level in Groundwater Using the Associative Rule Mining Algorithm with Random Forest Classification Approach

R. Siddthan[1], PM. Shanthi[2]

[1,2]*Department of Computer Science, J.J. College of Arts & Science,(Autonomous College) (Affiliated to Bharathidasan University), Pudukkotai, Tamil Nadu, India.*

Corresponding Author: sithan314@gmail.com

Received: 08 June 2023 Revised: 28 August 2023 Accepted: 15 December 2023 Published: 07 January 2024

Abstract - *In the past few years, the prediction of concentration in groundwater has received top emphasis in research on water resource management and pollution control. Therefore, the objective of the current research is to employ data mining techniques like the Random Forest (RF) methodology to determine the susceptibility zones of coastal districts in eastern India. Utilizing multi-collinearity analysis, fifteen conditioning parameters have been determined, and the association rule mining approach was used to determine the relative importance in order to create a groundwater concentration susceptibility map. To prepare the inventory dataset and related modeling purposes, the four K-Fold Cross Validation (CV) technique's resampling approach was applied. For assessing the effectiveness of all utilized models, seven statistical methodologies comprising receiver operating characteristics-area under curve (ROC-AUC) were employed. The study's findings indicated that boosting is the methodology that performs best for defining groundwater concentration susceptibility maps (GNCSMs) at the regional level. The results guarantee that the RF model is more effective compared to the boosting and bagging approach.*

Keywords *- Nitrate, RF, Performance, Prediction, Groundwater, Associative Rule Mining, Random Forest Classification.*

1. Introduction

One of the most widely available sources of pure water in the globe is groundwater. It can be used for a variety of domestic tasks, including drinking, manufacturing, irrigation, and other functions, but lately, overuse and limited water supplies have made it scarce. It is a reality that a significant portion of the global population, at least one-third of them, depends on groundwater for their daily needs in addition to drinking water. Groundwater availability has dramatically declined over the past few decades in tandem with population growth, and eventually, human activity has also contributed to declining water quality (Tyagi et al., 2013). According to a study, 780 million people worldwide do not have access to clean drinking water. However, groundwater contamination is now a significant barrier to regional sustainability and ecological stability (Güler et al., 2013; Wen et al., 2019). Due to agronomic practices, coastal groundwater is currently extremely vulnerable to metal pollution, nitrate contamination, salt intrusion, overexploitation, and contamination (A. R. Md. T. Islam et al., 2021). As the most oxidized chemical form of nitrogen in the nitrogen cycle, nitrate is a global problem due to groundwater pollution (Burow et al., 2010; Re et al., 2017). Nitrate is also a component of the nitrogen cycle. Due to its frequent occurrence in various countries across the world, including Africa (Talma and A, 2006), America (Power and Schepers, 1989), Europe (Strebel et al., 1989), Australia (Thorburn et al., 2003), and several developing Asian countries, this phenomenon has recently attracted a lot of attention (Chica-Olmo et al., 2014). Through rainwater infiltration, applied irrigation, excessive use of nitrogen fertilizer minerals in farming, and diffusion of poor-quality groundwater in several regions, this contaminant moves towards the groundwater zone. It poses an immediate risk to human health (Katsoulos et al., 2015). The United States Geological Survey (USGS, 2000) states that contamination resulting from water and land use can either be confined (point source) or widespread (non-point source). In addition to agricultural practices, other factors contributing to groundwater nitrate concentration include animal dung, industrial effluent, and ineffective sewage systems (Hansen et al., 2017). The "cumulative effect" of animal leftovers used as fertilizer, which significantly seep towards subsoil through irrigation or rainfall and contaminate the shallow aquifer groundwater, is discussed by Baker (1992) and Liu et al. (2005). Because coastal locations are closer to sea level than central areas, which are at mid- and high altitudes, the groundwater level there is shallower (Chang, 2014; Hagedorn et al., 2011). Thus, in coastal India, nitrate

contamination has been steadily rising daily, and this is the only supply of drinking water in the area (Khan et al., 2021; Mondal et al., 2008; Saranya et al., 2011). As per the findings of Townsend et al. (2003), nitrate has been identified as the predominant contaminant in groundwater since 1970. The EU Council Directive (91/676/EEC) safeguards waters against pollution caused by agricultural nitrate. The WHO (2017) recommended that the tolerable threshold value of nitrate in groundwater be fixed at 50 mg/l. If this value is exceeded, there are several negative effects on human health, including methemoglobinemia, which is commonly known as "blue baby syndrome" in infants.

Furthermore, it is a known fact that a nitrate concentration of less than 10 mg/l is deemed acceptable for human consumption. For this reason, measuring the nitrate content of underground drinkable water is essential for both supply and sustainable water management (Hansen et al., 2017; Voutchkova et al., 2021). According to Lawson (2011), one can raise their standard of living by enhancing the availability of safe water and the quality of their drinking water.

Groundwater is generally one of the significant sources of the populations' daily water requirements; however, it is contaminated by pollutants, including such nitrate, that penetrate through into the soil with water. Groundwater vulnerabilities and contaminants are serious concerns, particularly in densely populated regions, and necessitate careful consideration. In addition, to provide suggestions for agricultural systems movements, essential variables abstracted from remotely sensed Normalized Difference Vegetation Index time series (NDVI) have always been added to the database as an embedded technique, the feature significance acquired from RF, as well as CART, was implemented [1].

Whereas measuring the uncertainty of algorithms employed to estimate nitrate pollution from groundwater seems critical in groundwater conservation, it has primarily remained unrecognized. This problem prompts this research to investigate the prediction uncertainty of Machine Learning (ML) methods in this domain of investigation employing two alternative residual uncertainty approaches: Quantile Regression (QR) as well as uncertainty estimation relying on local errors as well as clustering. Artificial Neural Networks (ANN) and Support Vector Machines (SVM) were examined for their usefulness in forecasting contamination concentrations. The performance of the models is measured using a variety of indicators, both sensitive and insensitive to accuracy. Finally, the research assesses the relevance of the features by employing Shapley values to continuously rank features and provide model interpretability [2].

Assessing the uncertainty of ML approaches deployed to spatially simulate groundwater-nitrate pollution allows managers to make better risk-based decisions, increasing the reliability as well as the credibility of groundwater-nitrate predictions [3]. For modeling and validation, 109 nitrate concentration data points were employed. The efficacy of the four techniques was quantified using the ROC-AUC Curve. The results also demonstrate that integrating eight new components to the DRASTIC (Depth to water, net Recharge, Aquifer media, Soil media, Topography, Impact of vadose zone and Hydraulic Conductivity) improved the predictability of the Weights-of-Evidence (WOE) model, as the AUC value improved to 0.91. Gross replenishment is the utmost powerful influencing factor to groundwater vulnerability in the research area [4]. The aims and objectives are to (1) analyze the effectiveness of two Artificial Intelligence (AI) methodologies, notably ANN and SVM, in nitrate concentration modeling in groundwater using sparse data and (2) examine the consequence of data clustering as the pre-modeling strategy on the effectiveness of the advanced configurations [5].

The nitrate contamination map was developed using data of nitrate concentration from prior research. To assess the likelihood of groundwater contamination, three Machine Learning (ML) approaches have been utilized: SVM, MDA as well as Boosted Regression Trees (BRT). Furthermore, employing the ensemble modeling methodology, groundwater contamination probability maps have been generated. The models were validated and calibrated employing the AUC approach, with a minimal AUC threshold of around 80% attained. The algorithms' accuracy has been estimated to be in the 0.83-0.87 level [6].

Due to the scarcity of hydrogeological data, researchers are applying mathematical methodologies to increase the robustness of current techniques for evaluating the quantitative susceptibility of groundwater. The hybrid PSO-GA approach is a successful optimization algorithm that combines the benefits of Particle Swarm Optimization (PSO) as well as Genetic Algorithm (GA) while minimizing their drawbacks. The PSO-GA optimization technique is used to optimize the DRASTIC weighting system [7].

However, the instability of this measure is due to intrinsic flaws such as weight and rating assignment bias. The Stepwise Weight Assessment Ratio Analysis has been suggested as a novel DRASTIC adjustment utilizing a newly established Multi-Criteria Decision-Making (MCDM) strategy to modify the rating range; additionally, the Entropy, as well as GA methods, were used to change the relative weights of DRASTIC parameters [6, 8, 9]. Outputs from formerly physical-relied Central Valley models were used as predictor variables in the novel technique. According to three-dimensional visualization, nitrate forecasts are dependent on the likelihood of anoxic circumstances and other parameters, and nitrate estimates typically diminish with increased groundwater age [21-25].

The DRASTIC approach was used to develop a groundwater vulnerability map. The nitrate contamination map had been developed using nitrate concentration data from a prior investigation. To assess the likelihood of groundwater contamination, three Machine Learning (ML) models were used: SVM, MDA and BRT. Additionally,

using the ensemble modeling methodology, groundwater contamination probability maps were created. The models' accuracy was determined to being at the level of 0.82-0.87 [10, 26-29].

The contribution of the paper includes using the associative rule mining algorithm with the RF classification approach for nitrate level prediction in groundwater. The paper is organized as follows: Section 2 explains the related survey, Section 3 shows the proposed methodology, Section 4 shows the outcomes and discussions, and the paper concludes in Section 5.

2. Related Works

To create data sets for the validation of the RF model, several environmental parameters were compiled using remote sensing and Geographic Information System (GIS) approaches. The model performance was assessed using various parameters. These measurements show that the RF model for predicting groundwater was successful. The elevation of the water level was found to have the greatest relative influence on groundwater, according to the model's calculations of the relative importance of the predictor variables. He et al.'s [11] methodology offers a method for integrating many environmental elements into groundwater quality studies, which is important for long-term groundwater management in the Yinchuan Region. The development of an ANN model for the prediction of content in groundwater has been attempted by Wagh et al. [12]. The research region is located between latitude 19°55':20°25'N and longitude 73°55:74°15'E. One of the Godavari's tributaries, the river Kadava, rises in the Sahyadri Hills and flows from the northwest to the southeast. According to physicochemical findings, in both seasons, 67.50% and 75% of groundwater samples had NO_3 concentrations that were higher than the Bureau of Indian Standards (BIS) permitted limit (>45 mg/L). To model groundwater contamination, Alkindi et al. [13] used Bayesian approaches like the Bayesian generalized linear model (BGLM). Predictive modelling has used eleven conditioning factors as input parameters. The findings demonstrated that all of the Bayesian models utilized in this investigation were capable of simulating groundwater, with the BART model having the highest efficiency ($R^2 = 0.83$).

To create the first maps of groundwater pollution, Javidan et al. [14] investigated three multilayer Markov random-fields models. As a starting point model, RF was also applied. The performance characteristics of the models were evaluated using a number of cut-off-independent and cut-off-dependent evaluation criteria. In terms of producing maps of groundwater contamination, validation findings demonstrate that the conditional mixed MRF performed better than the other models. The main cause of contamination of groundwater is now agricultural operations. El Amri et al. [15] analyze the nitrate concentrations in a shallow aquifer, pinpoint the causes that can be attributed to them, and forecast future levels using Autoregressive Integrated Moving Averages (ARIMA) and ANN models. The findings revealed that levels in the pumping well are located throughout the plain, ranging from$17\ to\ 521\ mg\ L^{-1}$. 67% of the monitoring sites in total are significantly over the 50 mg L^{-1} as per the standard guideline threshold. Positive correlations exist between groundwater concentration and the major significant natural parameters, including land, soil texture, fertilizer application rates, groundwater table and livestock waste disposal. The ANN model demonstrated a good fitting among measured and simulated outcomes.

In order to estimate the content in the distribution system, Jamei et al. [16] created an accurate hybrid Boruta RF- Whale Optimization Algorithm (WOA) combined with an ANN. In order to assess the robustness of the WOA-ANN model for pattern prediction, kernel functions and multiple training strategies are applied as standalone validation models in combination with Support Vector Regression (SVR) and ANN techniques. The technique makes use of 11 variables that were taken from the experimental investigation and organized optimally as combinations of 5 input variables using regression and BRF Feature Selection analyses. The WOA-ANN model can be best optimized using the BRF-FS, according to the statistical and diagnostic assessments. With the best metrics ($R = 0.962, RMSE = 0.029\ mg/L, MAE = 0.024, and\ U = 0.056$) and a 30% increase in accuracy over the ANN, the suggested method was successful.

According to Di Nunno et al. [17], Nonlinear Autoregressive Neural Networks with exogenous inputs can produce precise models to forecast + nitrite concentrations in rivers. The Raccoon River and the Susquehanna River in the United States were used as case studies. Exogenous inputs included water temperature, water discharge, and specific conductance and dissolved oxygen. It was also determined how sensitive the forecasting was to variations in the time series length and exogenous input parameters. In order to assess the vulnerability in contamination of groundwater at a -contaminated area, Elzain et al. [18] compare the three ANFIS (the adaptive neuro-fuzzy inference system) with methodologies for evolutionary heuristics optimization like differential evolution algorithm (DE), GA, and PSO.

The South Korean city of Miryang was chosen for the study because it had both rural and urban functions and had a high level of contamination. To provide a sustainable and hospitable green environment, Hmoud et al. [19] devised an effective operation for monitoring drinking water. In this paper, the Water Quality Index (WQI) was predicted using the ANFIS algorithm. Feed Forward Neural Networks (FFNN) and K-nearest neighbours were used to categorize water quality. Although the dataset contains eight important parameters, only seven were found to have significant values. These statistical factors were used to build the suggested methodology. According to prediction findings, the ANFIS model was superior for predicting WQI values. However, for classifying water quality, the FFNN algorithm had the best accuracy (100%) (WQC).

Additionally, the FFNN model demonstrated higher resilience in classifying the WQC, whereas the ANFIS model correctly predicted WQI. In Andalusia, Spain, Cardenas-Martinez et al. [20] examined the effectiveness of the RF ML algorithm in conjunction with pollution prediction Feature Selection (FS) methods in Nitrate Vulnerable Zone (NVZ) groundwater sources over time and using updated environmental features. In an effort to make this methodology transferable to other places, a set of forty-four variables that are not intrinsic to groundwater bodies are to serve as the forecasters of the surroundings. The analysis of seasonal and inter-annual fluctuations in pollution also included additional dynamic variables arising from weather and livestock effluents. The predictive modelling evaluated changes in the likelihood that groundwater content would exceed 50 mg/L.

Subodh Chandra Pal et al. Numerous factors, such as wastewater, human activity, agriculture, and complex geo-hydrological parameters, contribute to the concentration of nitrate in subsurface water. Nitrate is also widely present in the shallow aquifer in coastal locations. For this reason, GNCSM is a useful strategy for managing the groundwater supply in coastal areas. In this work, forecast the nitrate concentration in groundwater in coastal locations using potential data-mining algorithms, such as boosting, bagging, and RF.

K.M. Ransom et al. In many places in the United States, groundwater is a significant supply of drinking water, and nitrate contamination is a cause for concern. Extreme gradient boosting, or XGB, is a three-dimensional machine learning model that was used to forecast nitrate concentrations for home and public supply zones throughout the continental United States (CONUS). Each zone has a different depth. The model correctly reproduces the distribution of low (≤1 mg/L) and high values (>10 mg/L) for hydrogeologic regions and predicts nitrate concentrations at the national scale (training R2 was 0.83 and hold-out R2 was 0.49). High nitrate concentrations affect only a small percentage (~1%) of the CONUS overall, with exceedances primarily seen in the Interior. Furthermore, high nitrate concentrations in the Interior Permian Secondary Hydrogeologic Region are identified by the model, which was not seen in earlier national studies. This work demonstrates that XGB is a valuable instrument for mapping groundwater quality at both regional and continental dimensions.

The drivers of anticipated nitrate concentrations at national and regional sizes were identified using the recently established SHAP technique (Lundberg and Lee, 2017). Nationally, well depth, climate, soils, and hydrology are the main factors influencing nitrate concentrations, which are often less than 1 mg/L. Nominal sources of nitrogen, other from land use proxy, were included in the top ten. The Piedmont and Blue Ridge carbonate-rock (PBRC) Principal Aquifer's high nitrate grid cells were subjected to SHAP study at the regional level. The main causes in the PBRC where the expected nitrate is expected to be higher than 10 mg/L were agricultural sources of nitrogen and elements that facilitate nitrogen transport. Analysis of SHAP data shows that black-box machine learning models are not necessary.

Aayush Bhattarai et al. Nine distinct machine learning algorithms were tested for their ability to predict nitrate and total phosphorus concentration for various types of watersheds: LR, F-SVM, M-SVM, kNN, RF, ANN, RT-BO, ensemble-BO, and GPR-BO. To begin with, the C-Q relationship for each watershed was examined in order to see how the kind of watershed affected the nutrient concentration prediction. In the urban Cuyahoga watershed, the nitrate concentration rose as stream flow increased, whereas in the rural and woodland watersheds, it diluted. Similarly, regardless of the type of watershed, the overall phosphorus concentration rose with stream flow.

The land-use distribution had an impact on the model performance for all strategies when it came to nitrate concentration prediction. Because nitrate concentrations in urban watersheds are regular and predictable, stream flow and month of the year are used as independent variables in more accurate modelling. Similarly, compared to the forested Grand watershed, ML models were more accurate in predicting the nitrate concentration in the agricultural watershed (Maumee, Raisin, and Sandusky). When it came to the $R2$ for the agricultural and urban watersheds, the ANN fared better than other ML models. Conversely, RT-BO performed better than other ML models for the wooded Grand watershed.

Similarly, for every kind of watershed, the Bayesian optimized RT, ensemble, and GPR continuously produced good results. Increasing stream flow in agricultural and wooded watersheds may cause more soil erosion, which in turn may raise the concentration of particulate phosphorus in the water column. Thus, with stream flow, total suspended solids, and month of the year as independent variables, the model predictability was higher for agricultural and forested watersheds in predicting phosphorus concentration. On the other hand, point sources like wastewater treatment facilities are the primary source of phosphorus inputs in an urban watershed. As a result, the model's predictability of phosphorus content was somewhat compromised. In terms of the $R2$ for the test data, the ANN seems to perform better than other ML models for the urban Cuyahoga watershed, while the produced ML models underfitted the training dataset. In contrast, when it came to forecasting the total phosphorus concentration for the agricultural and wooded watershed, ensemble-BO and M-SVM fared better than other ML models.

2.1. Review Literature

The solution to food production in the early 20th century to meet the demands of the world's rapidly expanding population had already been shown to be a significant threat to water resources by the late 20th century. The intense use of fertilizers in agriculture is mostly to blame for the paradoxical truth that, in order to meet the

demand for food, we are severely stressing groundwater—the most important and valuable drinking water resource—in the process. Global agricultural output increased dramatically as a result of the Haber-Bosch process's discovery, which made it easier to produce fertilizer industrially (Erisman et al., 2008). Interference with the nitrogen cycle can have detrimental effects on human health as well as extensive negative environmental effects, including water pollution (Erisman et al., 2013; van Grinsven et al., 2006; Vitousek et al., 1997). The EU Water Framework Directive 2000/60/EC was put into effect as a result of several water-related directives being passed in the middle of the 1970s when water contamination started to become a public concern (Kallis and Butler, 2001). In order to provide integrated protection and preserve or return water systems to their "good status," the Water Framework Directive was developed. The Nitrates Directive 91/676/EEC, which is a crucial component of the Water Framework Directive, serves as a safeguard for water bodies against the demands of agriculture. Effective mitigation strategies are needed to lower nitrogen inputs to surface and groundwater in order to fulfill the high requirements set by the Water Framework Directive (WHO's maximum contamination level: 50 mg NO -/l or 10 mg NO -N/l in drinking water).

3. Proposed Methodology

3.1. Machine Learning Algorithms

The standard technique for estimating the concentrations of phosphorus and nitrate in all watersheds is Linear Regression (LR).

3.1.1. Linear Regression (LR)

LR uses a linear equation, or a first-order polynomial equation, to fit the data. In the case of data with linear relations, this approach is helpful. Mathematically, the standard linear regression model looks like this:

$$\hat{Y}=\sum (aX+b) \quad (1)$$

Where the input and output variables are denoted, respectively, by X and $\hat{Y}$.Using the actual and anticipated data, the least-squares approach is used to fit the model coefficients (a and b).

3.1.2. k Nearest Neighbors (kNN)

kNN Models rely on the proximity of data points wherein characteristics and training patterns are used to classify a new object. In this case, the average or the designated number of nearby values is the output or the predicted value. The value k is used to characterize the given number. The formula for kNN is:

- Initialize k;
- Determine the query example's Euclidean distance from the labeled examples.

$$d(x,x') = \sqrt{((x_1-x_1')^2 +(x_2-x_2')^2 + \ldots\ldots\ldots+ (x_n-x_n')^2)} \quad (2)$$

Where, (x,x') is the sample point;

- Sort the labeled instances in order of smallest to largest
- Use cross-validation to determine the ideal number k of nearest neighbors based on RMSE
- Use kNN to compute an inverse weighted average.

3.1.3. Regression Tree (RT)

Recursive partitioning is the foundation of an RT method for nonlinear regression. The practice of dividing the data into more manageable segments and then dividing each partition into even smaller sub-divisions is known as recursive partitioning. Recursive partitioning is represented by RT as a tree, where each terminal node is a partition cell. Although RT is easy to model and see, ensembles were created because a single-tree model was unstable.

3.1.4. Ensemble

To create a more accurate RT model, ensemble approaches integrate a number of weak RT models. The ensemble uses distinct samples from the original data set and combines their results to produce numerous diverse regression models. In this work, two different Ensemble methods are used: bagging and least-squares boosting or LSBoost.

Bagging uses random sampling with replacement to produce a large number of training sets. Every data set is subjected to the RT method, and the models' average is then calculated to determine the predictions for the data that have not yet been observed.

In Boosting, records with poor prediction in earlier models are the focus of successive model pieces of training, which produces a more accurate model. After completion, a weighted majority vote combines all of the predictors. The difference between the observed response and the total prediction of all learners that the ensemble has previously grown in LSBoost is used to fit a new learner.

3.1.5. Random Forest (RF)

An RF is another kind of ensemble that was used in this research. By using a majority vote, RF creates several decision trees that are then used to categorize fresh instances. Every decision tree node uses a portion of the original set of randomly chosen attributes. Likewise, every tree, including bagging, uses a distinct bootstrap sample data. There is a possibility that RF will generate hundreds or maybe thousands of trees. Ten trees are chosen for this assignment in order to create a forest.

3.1.6. Artificial Neural Network (ANN)

An ANN solves a difficult problem by utilizing connected units between the input and output layers. In the current work, MLP is one of the ANN topologies used. It differs greatly from polynomial regression in that the output layer employs an activation unit, and the connected unit is typically a simple linear equation. Similar to a logical switch, the activation unit is only turned on when certain threshold values are reached. Typical categories of activation functions include Sigmoid and ReLu. This study uses ReLu activation.

Typical MLP with a hidden layer can be mathematically modeled as follows:

$$y_i = \sum_j^n w_{i,j} x_j + b \quad (3)$$

Where *x* is the input, *w* is the weight, and *b* is the bias in the hidden layer.

3.1.7. Support Vector Machine (SVM)

Complex relations, which are difficult for lower-order polynomial equations to adequately express, can be modeled with the help of SVM-based regression models. Pattern recognition, classification, regression, and prediction issues are all addressed by SVM, a potent supervised learning technique with outstanding generalization capabilities. The equation is used to determine the expected value:

$$\hat{Y} = \sum_{i=1}^{n} K(X_i, X_0)(\alpha_i - \alpha_i^*) \quad (4)$$

Where αi and $\alpha * i *$ are the kernel function, and (Xi, $X0$) are the support vectors. To apply its regression learner, the SVM function can be combined with several kernel functions (KF). The Gaussian Kernel Function (GKF), which is defined as follows, is frequently applied to SVM classification and regression.

$$K(X_i, X_0) = \exp\left(-\|x_i - x_j\|^2\right) / \left(2\sigma^2\right) \quad (5)$$

Both medium and fine Gaussian SVMs (F- and M-SVMs) are employed in this work. The definition of medium and fine Gaussian depends on how thin the Gaussian function being applied is.

The objective of the current research is to employ data mining techniques like the Random Forest (RF) methodology to determine the susceptibility zones of coastal districts in eastern India. The fifteen conditioning parameters have been determined, and the association rule mining approach was used to determine the relative importance in order to create a groundwater concentration susceptibility map. The workflow is shown in Fig.1

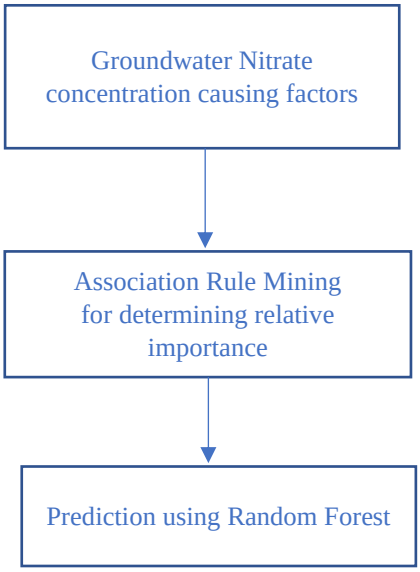

Fig. 1 Workflow of the proposed method

3.2. Spatial Association Rule Mining

Applying association rule mining (ARM) to spatial datasets is known as spatial association rule mining. A geographic association rule is in the manner of $P1 \wedge P2 \wedge \ldots \wedge Pm \rightarrow Q1 \wedge Q2 \wedge \ldots \wedge Qn$ (sup% and con%). It indicates a relationship of association between a group of predicates Pa $(a = 1, \ldots, m)$ and Qj $(j = 1, \ldots, n)$, each of which contains at least one spatial predicate. Indicating a spatial orientation or representing topological relationships between spatial objects (such as intersecting or containing) are two examples of spatial predicates (e.g., north, left). The number of transactions that involve both the consequent and the antecedent of the rule is measured by the support of the rule (sup%). The number of transactions that satisfy the consequent and the antecedent of the rule are indicated by the confidence of the rule (con%).

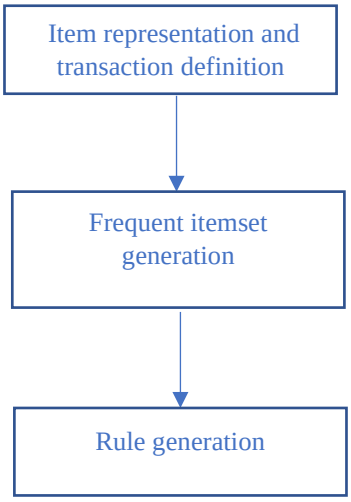

Fig. 2 Steps in associative rule mining

1. Fix “items” as well as “transactions” for spatial datasets.
2. Describe all the itemsets that meet the minimal support criterion in the frequent item set generating step.
3. Create rules from often occurring itemsets that meet the minimal confidence requirement.

In Subtasks 2 and 3, apriori-style association mining techniques are frequently employed. Objects must be described by categorical qualities in order for these algorithms to function. Continuous characteristics must be discretized in Subtask spatial space as a result. It uses the same transaction model in our work. The enormous number of created patterns presents a difficult challenge for spatial ARM, particularly in real-world uses. Many relationships are either explicitly reflected in geographic databases or already have established geographic dependencies. For instance, it is a well-known and uninteresting relationship that petrol stations are typically located.

3.3. Data Collection and Data Pre-processing

The Texas Water Development Board, the state organization in charge of statewide water planning, maintains the Texas Ground Water Database (GWDB), from which the datasets used in this study were taken. Over

the past 30 years, the GWDB has tracked and examined Nitrate content. In extremely high concentrations, it is toxic. Even at low levels, prolonged exposure to Nitrate can raise the risk of developing cancer. Nitrate comes from both natural and artificial sources, including mine drainage, hydrothermal leaching of Nitrate-containing minerals or rocks, mine tailings, insecticides, and biocides. The World Health Organization has identified Nitrate as one of the major criteria for assessing the quality and safety of drinking water in the United States, Mexico, Thailand, Hungary, India, Chile, Ghana, Bangladesh, China, and Argentina. Datasets must be cleaned in order to address issues like inconsistent data, missing values, and duplicate entries due to changes in GWDB's data collection and management standards and practices over time. For each water well, the resulting Nitrate spatial dataset contains class labels (CL), nonspatial characteristics (A), and spatial attributes (S). Some spatial information, including the zone, river basin, longitude, and latitude, are taken straight from the database. The intersection model is used to estimate implicit spatial features, like the separation between rivers and wells. With the help of subject matter specialists, nonspatial parameters, such as well depth, fluoride, and other chemical metal element concentrations, like iron, Vanadium, selenium, and molybdenum, are chosen. Molybdenum and Vanadium have comparable geochemical behaviour, and the attributes iron, selenium and fluoride may point to the ultimate source of Nitrate. Among these features, well depth is employed for studies on mobilizing mechanisms. Water wells are divided into two categories: safe and harmful. According to the Environmental Protection Agency's regulation for drinking water, a well is unsafe if its Nitrate concentration is greater than 10 g/l.

After data preparation, 11,922 records were chosen from the GWDB. The support for the Nitrate class attribute was enhanced by this discretization technique, making it easier to identify supervised association rules using Nitrate. As a result, the technique can successfully identify the supervised association rules for the various classes of Nitrate. It has been demonstrated that the strategy, which results in mismatched bin sizes, yields better outcomes in data mining tasks. Table 2 contains a list of each continuous attribute's splitting points. Table 1 contains a list of the nonspatial features used in the Nitrate dataset. One of two techniques—supervised discretization with class information or unsupervised discretization without class information—is frequently used to discretize continuous attributes.

Table. 1 List of the nonspatial features used in the nitrate dataset

Total No. of wells	11,022		
Nonspatial Attributes	Mean	STD	Splitting Points
1. well depth	567.945	633.902	214.9
2. (mg/l)	11.289	29.037	0.083, 0.387, 15.9, 27.4

3.4. Regional Association Rule Mining

For each of the detected regions, frequent itemsets are generated using the Supervised Apriori Gen technique. For the experiments, we set minimum support and confidence limits of 10% and 70%, respectively.

A = $\{a_1, a_2, \ldots \ldots a_n\}$ is a set of nonspatial attributes, CL = $\{cl_1, cl_2, \ldots \ldots cl_n\}$ is a set of class labels, and let D be a spatial dataset. S = $\{s_1, s_2, \ldots \ldots s_l\}$ is a set of spatial attributes.

$$I = S \cup A \cup CL \tag{6}$$

$$I = \{s_1, s_2, \ldots \ldots s_l, a_1, a_2, \ldots \ldots a_n, cl_1, cl_2, \ldots \ldots cl_n\} \tag{7}$$

Nominal attributes are created from continuous attributes. Let T represent the set of all transactions, with t = $\{t_1, t_2, \ldots \ldots t_N\}$. A relational table with N tuples that adhere to schema I (which has l + m + n number of items) can be used to depict T. As a result, an item i ∈ I is a binary variable, and its value is 0 otherwise and 1 if the item is present in t_i $(i = 1, \ldots, N)$. As a result, the given class structure CL is used to classify the set of transactions T.

A supervised association rule r is of the form P → Q, where P ⊆ I, Q ⊆ I, and (P∪Q)∩ CL= Ø.

With confidence con, and support sup, the rule r is upheld in the D.

$$\text{sup}(P \rightarrow Q) = \frac{\sigma(P \cup Q)}{N} \tag{8}$$

$$\text{con}(P \rightarrow Q) = \frac{\sigma(P \cup Q)}{\sigma(P)} \tag{9}$$

The support count is defined as σ(α) = |{t_i|α ⊆t_i, t_i ∈ T}|. If a supervised association rule meets user-specified thresholds for minimum confidence (min_ confidence) and minimum support (min_ support), it is considered strong.

The objects that are part of a region R are indicated by the extension of R, or EXT(R). A region needs to be contiguous, meaning that there should always be a path connecting any two objects that are part of the same region. Our region discovery algorithm uses a reward-based evaluation scheme to assess the quality of the generated sub-regions. Given a global region R, a dataset D, where D = EXT(R), and an underlying class structure CL, the algorithm finds the best region. The sum of the rewards from each sub-region Ri(i = 1..m) defines the fitness function, which assesses the quality of the generated sub-regions. R_X = $\{R_1, \ldots . R_m\}$.

$$q(R_X) = \sum_{i=1}^{m} reward(R_i) = \sum_{i=1}^{m} (interest(R_i) \times (R_i)^{\beta}, where\ \beta > 1. \tag{10}$$

Subregions $R_1, \ldots . R_m$ are located so that:

1. The subregions are not connected: i = j, EXT(R_i) ∩EXT(R_j) = Ø.
2. $q(R_X)$ is maximized by R_X = $\{R_1, \ldots . R_m\}$.
3. The generated subregions, that is, EXT(R_1) ‖

EXT(R_m) ⊆ EXT(R) does not have to be exhaustive with respect to R.

4. The rewards that each region receives determine how $R_1, \ldots R_m$ are ranked. Subregions that yield little or no reward are often eliminated.

3.5. Multi-Collinearity Analysis

The relationship between two or more other multiple conditioning parameters together in a linear regression is known as multi-collinearity. This data-related issue has a negative impact on the study, and identifying it is crucial for all researchers since it may limit the model's potential to be generalized. The absence of multi-collinearity problems in this study should be ensured, and this can be quantified using a variety of methods, including the conditional index, variance decomposition proportions and tolerances (TOL). To minimize potential mistakes among the components taken into consideration in this concentration investigation, the statistical methods VIF and TOL have been used. The values of VIF and TOL have an inverse connection when these values are below 10 and above 0.1, respectively. Calculating TOL and VIF requires the use of the following equations:

$$\text{TOL} = 1\text{-}R^2 \tag{11}$$

$$\text{VIF}=1/\text{TOL} \tag{12}$$

3.6. Random Forest (RF)

RF is a well-liked ensemble data mining approach that works with a lot of decision trees for classification and regression. As a non-parametric supervised ML technique, it generates numerous sets of samples through sampling and performs multiple regression tree training stages; as a result, it estimates the classification of the data based on the voting results of various classifiers. Regression trees have severe issues with the outfitting of the datasets in training, which results in inadequate functioning by providing an unclear dataset. The RF is known to be the best solution to address these issues.

4. Results and Discussion

4.1. Performance Measures

A research project must include accuracy assessment and model validation measures. The produced outcome has no impact in actuality without validation. In order to validate the proposed models, seven statistical methodologies were used, including the specificity (SPE), sensitivity (SEN), accuracy (ACC), precession (PRE), Kappa index, F-score and AUC- ROC statistical method. To characterize the outcome result's quality, four index considerations, true positive (TP), false positive (FP), true negative (TN), and false negative (FN), were implemented. The ratio of true positives to all positives, which is a measure of precision, varies from 0 to 1; if the attributes seem to be significantly greater, the models produce accurate outcomes; the kappa index, which ranges from 1 (undependable) to +1 (dependable), whenever the value is 0, illustrates a poor correlation among estimation as well as observation in the framework. The framework categorizes kappa values into five classifications, including slight and very slight (0.1 to 1.0). AUC-ROC, which could accurately discriminate among occurrence and non-occurrence phenomena with values varying between 0 (inaccuracy) and 1, was an additional validation methodology employed in this work (consistent). Meanwhile, the lesser (0.5) and greater (1) values represent, respectively, poor and phenomenal performance. Therefore, the X and Y axes, respectively, show their effectiveness metrics.

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \tag{14}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{16}$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{17}$$

$$\text{F1 Score} = 2*\frac{\text{Precision}*\text{Recall}}{(\text{Precision}+\text{Recall})} \tag{18}$$

$$\text{Kappa} = \frac{P_a - P_{exp}}{1 - P_{exp}} \tag{19}$$

$$P_a = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{20}$$

$$P_{exp} = \frac{(\text{TP}+\text{FN})(\text{TP}+\text{FP})+(\text{FP}+\text{TN})(\text{FN}+\text{TN})}{\sqrt{(\text{TP}+\text{TN}+\text{FP}+\text{FN})}} \tag{21}$$

$$\text{AUC} = \frac{(\sum TP+\sum TN)}{(\text{P}+\text{N})} \tag{22}$$

The corresponding overall and randomized accuracy is expressed by P_a and P_{exp}.

4.2. Discussion of Results

In the aforementioned investigation, multi-collinearity analyses were conducted out in four folds to establish whether the variables were collinear before additional models were established. Following a positive multi-collinearity evaluation, fifteen nitrate conditioning parameters were identified. According to Table 2's multi-collinearity statistics, there existed a considerable multi-collinearity concern with folds 2, 3, 4, and 3. The tuning parameter, commonly referred to as the tuning hyper-parameter, is frequently employed for ensemble model evaluation. This methodology of hyper-parameter adjustment aids in the resolving of optimal learning classification issues. The best training dataset is employed for accurate assessment of learning ensemble strategies.

In this work, adjust characteristics of Bagging, Boosting as well as RF approaches were illustrated utilizing boosting repetitions and Root Mean Square Error (RMSE) assuming Max Tree Depth (MTD) are displayed in Fig.3, 4 & 5. The association between boosting iterations and RMSE was determined using the mean, median, and optimal value from boosting, relying on the three assumptions indicated above.

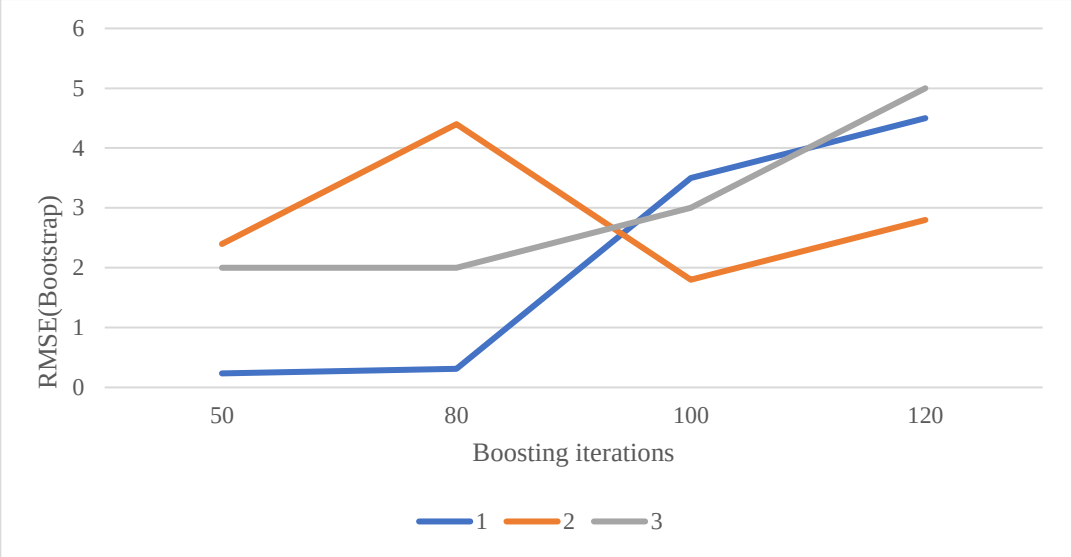

Fig. 3 Considering the MTD and model analysis with tuning values employing boosting

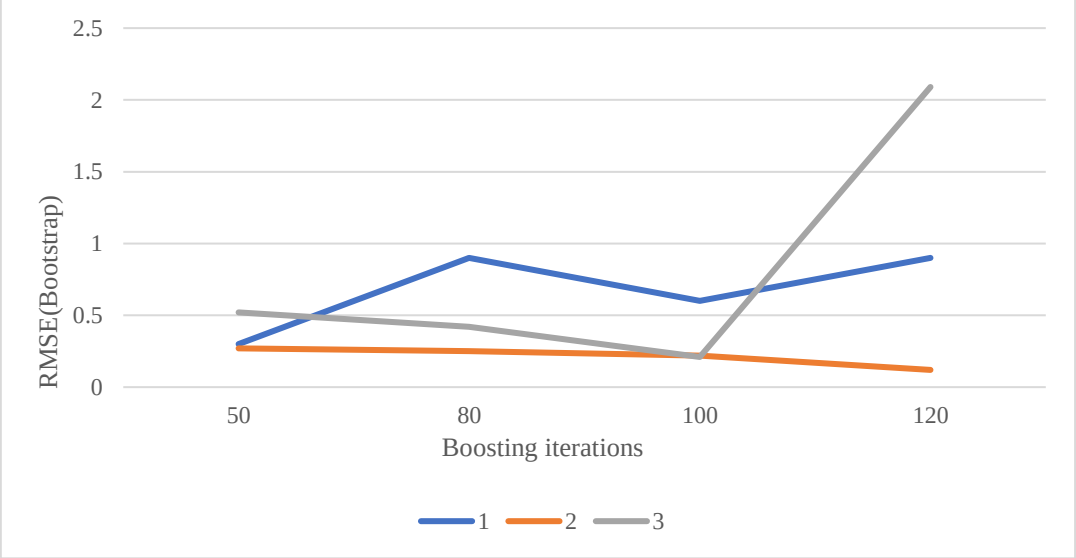

Fig. 4 Considering the MTD and analysis of designs employing bagging to tune the variables

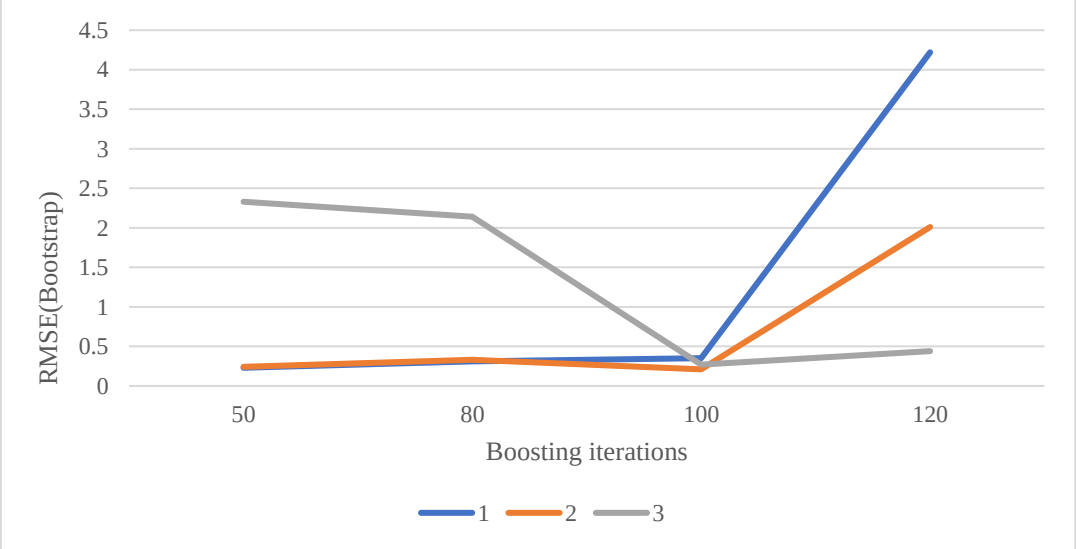

Fig. 5 Considering the MTD and analyzing the RF model with tuning variables

Table. 2 Variable importance value of conditioning factors

Factors	Fold 1	Fold 2	Fold 3	Fold 4
PO_4^{-}	0.754	0.689	0.764	0.889
Depth	0.714	0.695	0.689	0.739
Temp	0.532	0.562	0.579	0.574
SO_4^{2-}	0.421	0.446	0.516	0.452
K^+	0.252	0.276	0.328	0.315
Ca^{2+}	0.232	0.255	0.251	0.258
Na^+	0.213	0.239	0.232	0.237
EC	0.165	0.214	0.184	0.171
$HCO3^-$	0.154	0.173	0.156	0.164
Cl^-	0.144	0.154	0.143	0.148
Mg^{2+}	0.054	0.089	0.096	0.110
Salinity	0.113	0.124	0.135	0.124
pH	0.029	0.059	0.049	0.069
AS	0.009	0.019	0.019	0.039
F^-	0.953	0.854	0.863	0.965

The model evaluation outcomes of the overall six statistical measurements employed in this investigation which have been established by employing either training as well as testing datasets for overall K-Folds and evaluating three strategies. All indicators having larger values indicate the optimal performance of every algorithm. All of the outcomes in this study show nearly identical outcomes in the training as well as testing datasets for overall folds. With the exception of fold 1 in the training dataset, boosting outperforms bagging as well as RF in GNCSM delineation. It is understood that improving data mining models reduces time and aids in reducing extraneous information, whereas RF usually requires a more extensive dataset and a considerable amount of effort.

According to non- and high-dimensional difficulties, the other two algorithms, bagging and RF, produced greater appropriate outcomes in all four folds than boosting. The observations indicate that numerous areas of the research region are grievously impacted by nitrate concentration, with the very excessive concentration zone exhibiting the most vulnerability, accompanied by the higher, lower, and intermediate categories.

The RF ensemble methodology was implemented in the investigation to identify and prioritize the aforementioned fifteen nitrate causal elements (Table.2), and their respective relative relevance is illustrated in Fig.6.

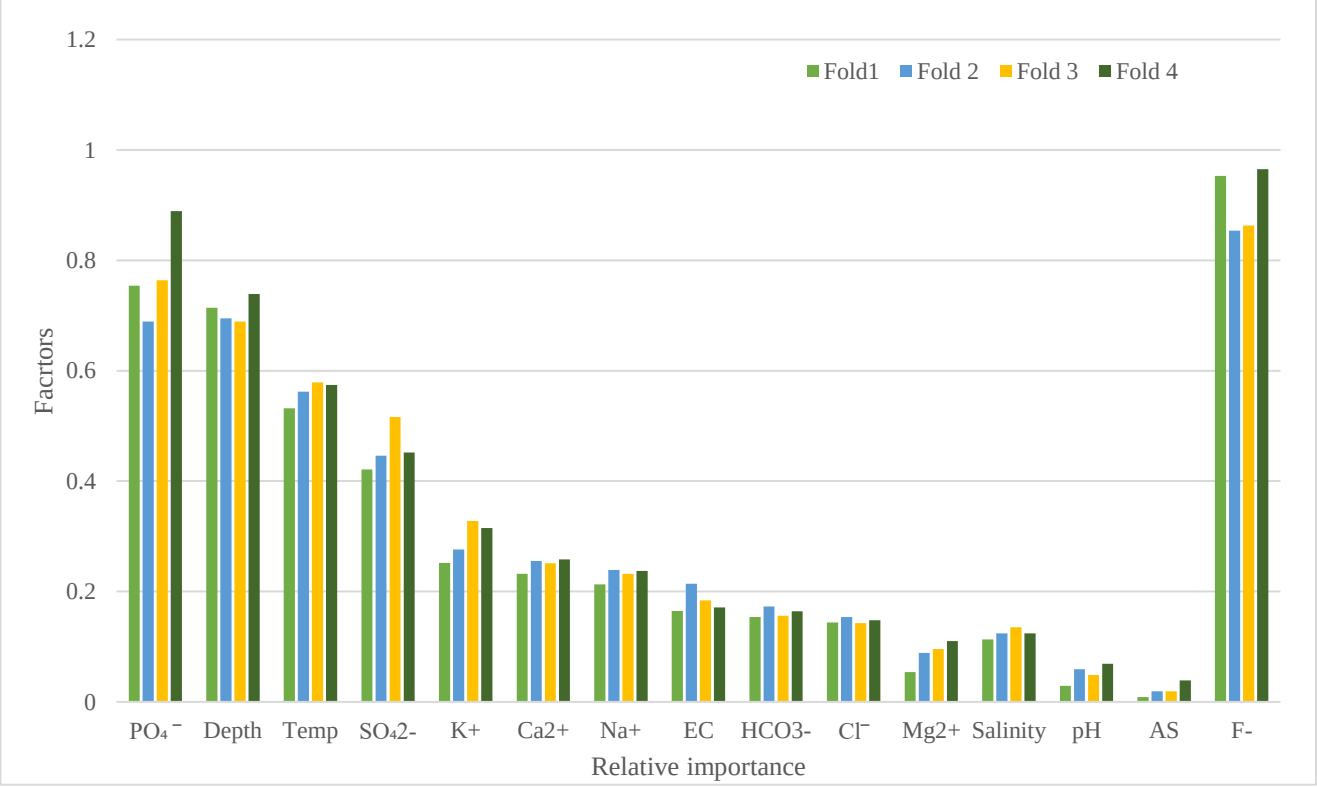

Fig. 6 Graphical presentation of relative importance value for Fold 1 to Fold 4.

Employing the four K-fold CV approach, the distributions of each susceptibility class's surface area. These are the relatively important values of three requirements and are ranked from 1 through 15 within this sequence, such as F^- , PO_4^{3-} and depth, dependent on whichever factors possess the most influence on nitrate concentration, ranked first, second, and third, respectively.

According to the Nitrate study literature, the first five rules that we offer for the regions examined are all significant and meaningful. In regions with large concentrations of Nitrate, mining regional rules reveal characteristics; in regions with low concentrations, it reveals characteristics.

$$(X, Well) \wedge (X, 0 - 0.085) \rightarrow Nitrate\ level(X, dangerous)\ (100\%).$$

The rule asserts with absolute certainty that Region 3 wells with concentrations less than $0.085\ mg/l$ contain harmful levels of Nitrate. Hudak's study in environmental geology confirms the substantial correlation of high Nitrate content.

$$(X, W\ ell) \wedge vanadium(X, 20.05 - 37.95) \wedge selenium(X, 74.55 - \infty) \rightarrow Nitrate\ level(X, dangerous)\ (100\%).$$

The rule specifies with absolute certainty that wells in Region 1 that have selenium concentrations greater than $74.55\ g/l$ and vanadium concentrations between 20.05 and $37.95\ g/l$ have unsafe Nitrate concentrations. Our experimental findings also reveal a few fresh rules that haven't been discussed in Nitrate analytical literature.

$$(X, W\ ell) \wedge depth(X, 0 - 215.5) \wedge iron(X, 19.65 - 20.05) \rightarrow Nitrate\ level(X, dangerous)\ (100\%).$$

According to the rule, shallow wells with iron concentrations in a particular range have higher Nitrate concentrations. We anticipate that the findings of our study will assist subject-matter specialists in choosing intriguing theories for additional scientific investigation. Additionally, it would be interesting to know if there are regional variations in the laws. The rule sets produced for Regions 1 and 3 (hot areas), as well as Regions 2 and 4, were compared (cold spots). The geographic risk patterns linked with Nitrate vary by region due to the studied area's varied topography and agricultural activity.

$$(X, W\ ell) \wedge (X, 28.085 - \infty) \wedge fluoride(X, 4.605 - \infty) \rightarrow Nitrate\ level(X, dangerous)\ (100\%).$$

In contrast, the nitrate contribution or otherwise prediction modeling technique is less affected by pH as well as salinity. As a result, the modeling of nitrate concentration takes into account these fifteen parameters in various degrees. For illustration, in the investigation for GNCSM, all fifteen characteristics have been included.

The model assessments across most selected statistical measures are represented in Tables 3 and 4. Fig. 7 illustrates the spatial variation of GNCSM throughout the proposed investigation area through the wetter period. These outcomes have been generated employing training as well as testing datasets across those K-Folds as well as three scenarios. The optimal effectiveness of every modeling is characterized by all measures with greater values. All of the outcomes in this study show nearly identical results in the testing as well as training datasets for all folds.

Table. 3 Models' tendency for prediction utilizing a nitrate training dataset

Stage	CV Fold	Model	Accuracy
Training stage	Fold 4	RF	0.946
		Bagging	0.965
		Boosting	0.973

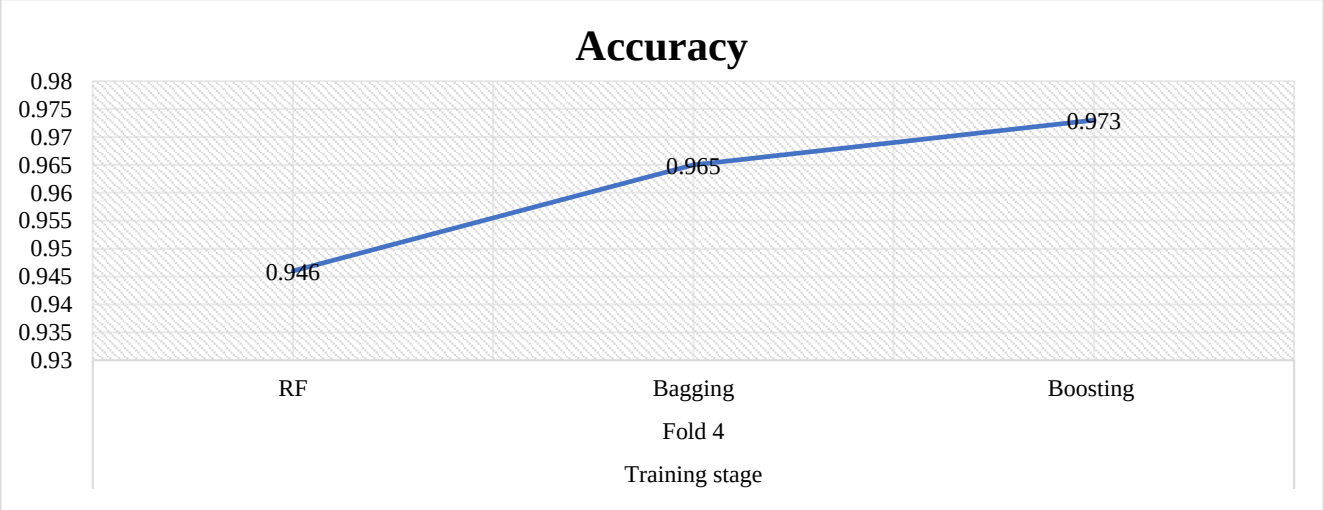

Fig. 7 Employing the prediction utilizing a nitrate training dataset

Table. 4 Models' capability for prediction with a nitrate test dataset

Stage	CV Fold	Model	Accuracy
Validation stage	Fold 4	RF	0.942
		Bagging	0.955
		Boosting	0.957

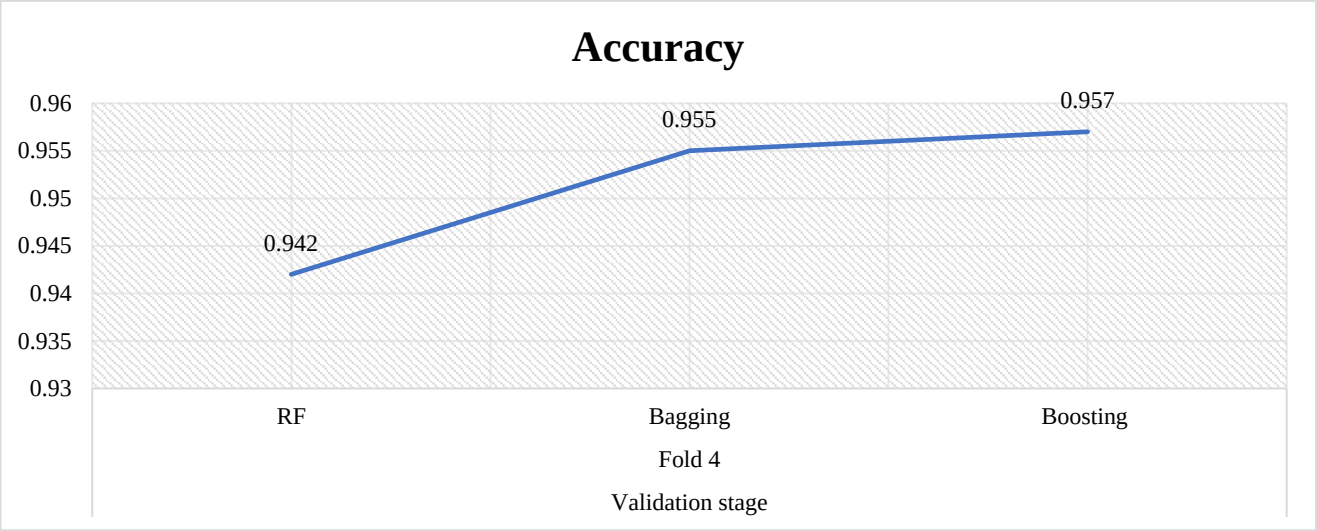

Fig. 8 Employing the prediction utilizing a nitrate validation dataset

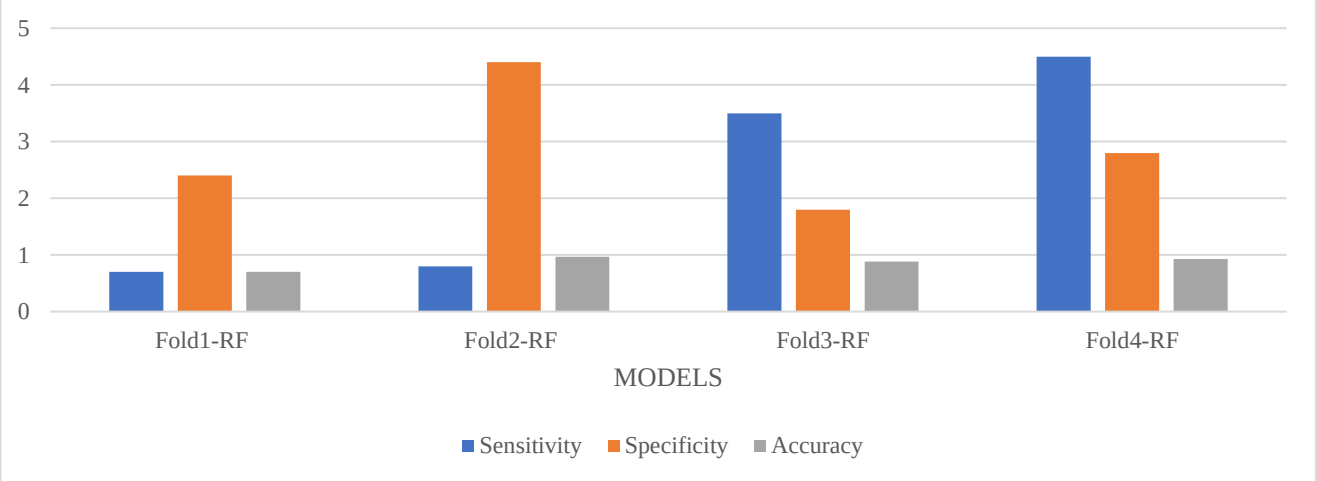

Fig. 9 Evaluation of models through Statistical Indices

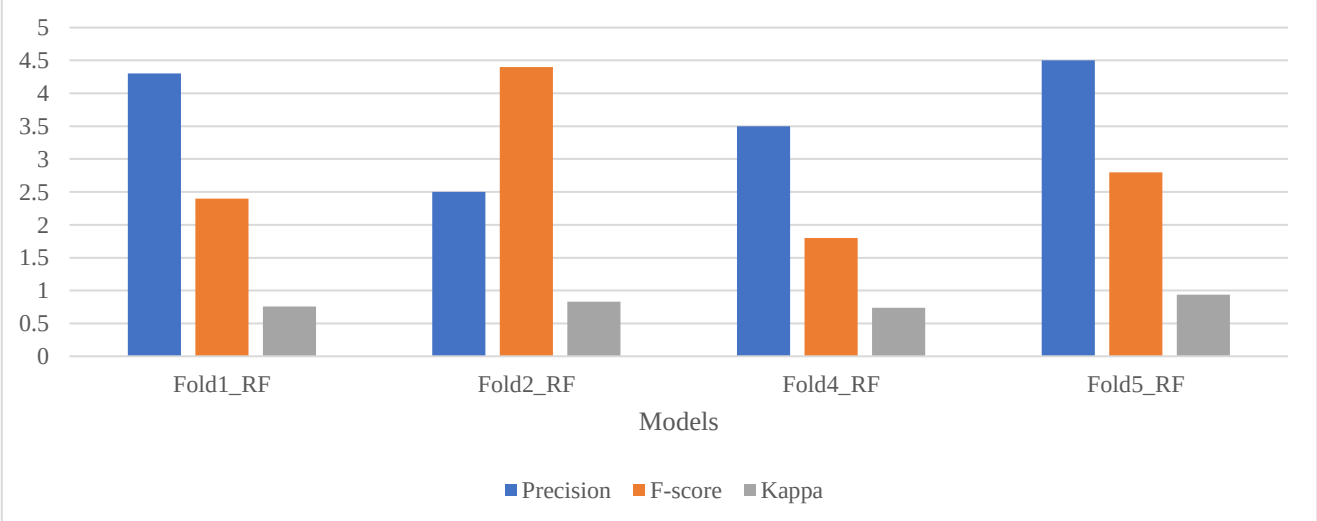

Fig. 10 Model evaluation through statistical indices

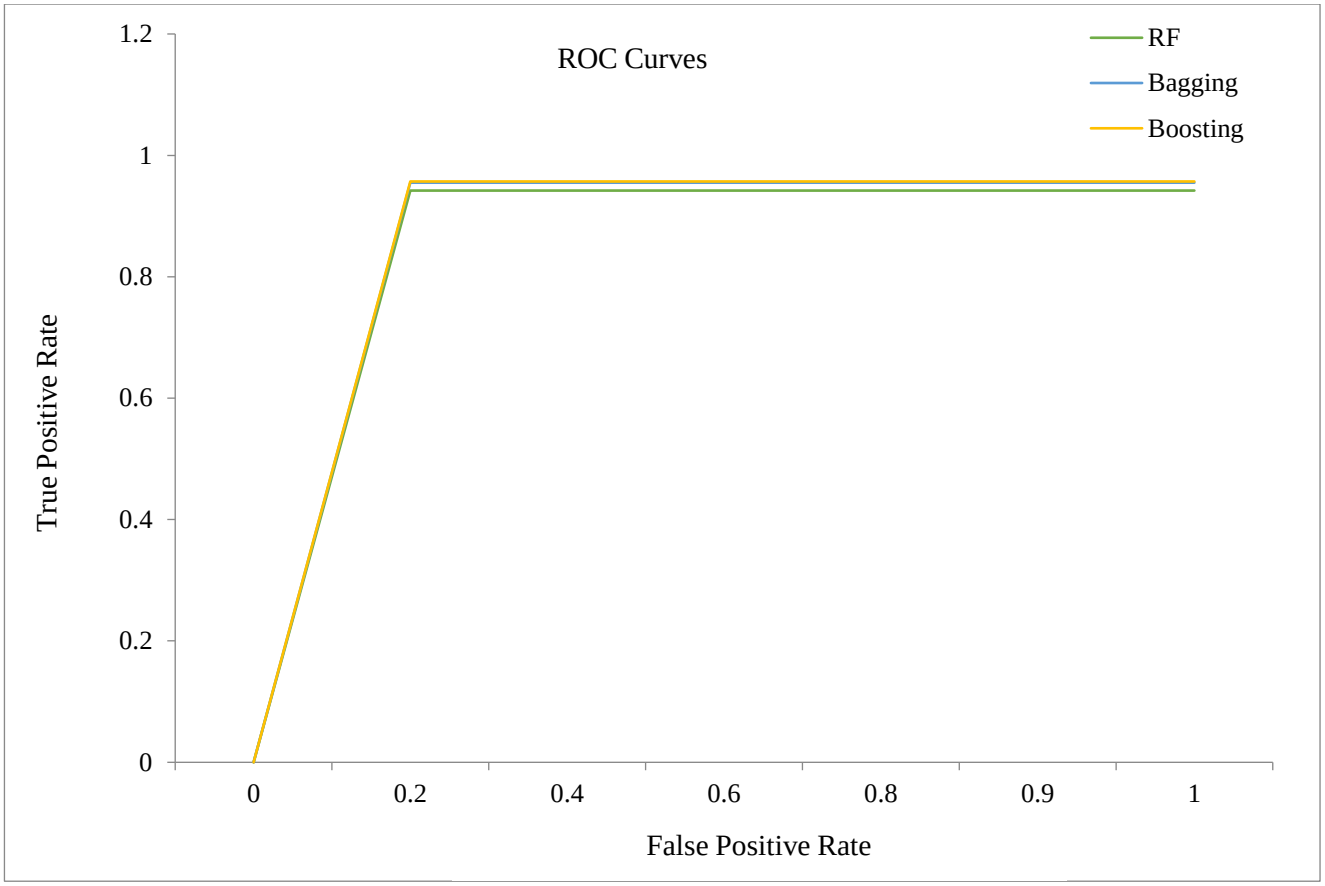

Fig. 11 ROC curve comparison for each model

The ROC curve is utilized to assess the separate modelling performances of GNCSM in eastern India's coastal regions. The outcomes demonstrate that Boosting produces dependable performance results later. Figure. 11 demonstrates the RF as well as bagging models. Boosting generates the most dependable results of the three methodologies tested, and it is the highest-performing model in both the training and testing datasets.

5. Conclusion

Nitrate concentrations in groundwater sources were generated from a variety of sources, comprising wastewater, human impacts, agricultural operations, and complex geo-hydrological characteristics, and nitrate is also plentiful in coastal shallow aquifers. Thus, the proposed investigation attempts to determine nitrate susceptibility zones in coastal districts of eastern India by employing three data mining strategies: RF, boosting, and bagging. The results guarantee that the RF approach is significantly effective. Because the farmers and residents rely on groundwater for both drinking and cultivation purposes, nitrate pollution exposes a large population to health risks. As a result, mitigation and continuous checking of the quality of the groundwater are critical over time.

References

[1] Victor F. Rodriguez-Galiano et al., "Feature Selection Approaches for Predictive Modelling of Groundwater Nitrate Pollution: An Evaluation of Filters, Embedded and Wrapper Methods," *Science of the Total Environment*, vol. 624, pp. 661-672, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Shine Bedi et al., "Comparative Evaluation of Machine Learning Models for Groundwater Quality Assessment," *Environmental Monitoring and Assessment*, vol. 192, pp. 1-23, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Omid Rahmati et al., "Predicting Uncertainty of Machine Learning Models for Modelling Nitrate Pollution of Groundwater Using Quantile Regression and UNEEC Methods," *Science of the Total Environment*, vol. 688, pp. 855-866, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Khabat Khosravi et al., "A Comparison Study of DRASTIC Methods with Various Objective Methods for Groundwater Vulnerability Assessment," *Science of the Total Environment*, vol. 642, pp. 1032-1049, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jawad S. Alagha, Md Azlin Md Said, and Yunes Mogheir, "Modeling of Nitrate Concentration in Groundwater Using Artificial Intelligence Approach—A Case Study of Gaza Coastal Aquifer," *Environmental Monitoring and Assessment*, vol. 186, pp. 35-45, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] Muhammad Awais et al., "Assessing Nitrate Contamination Risks in Groundwater: A Machine Learning Approach," *Applied Sciences*, vol. 11, no. 21, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Maryam Torkashvand et al., "New Hybrid Evolutionary Algorithm for Optimizing Index-Based Groundwater Vulnerability Assessment Method," *Journal of Hydrology*, vol. 598, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Maryam Torkashvand et al., "DRASTIC Framework Improvement Using Stepwise Weight Assessment Ratio Analysis (SWARA) and Combination of Genetic Algorithm and Entropy," *Environmental Science and Pollution Research*, vol. 28, pp. 46704-46724, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Katherine M. Ransom et al., "A Hybrid Machine Learning Model to Predict and Visualize Nitrate Concentration throughout the Central Valley Aquifer, California, USA," *Science of the Total Environment*, vol. 601-602, pp. 1160-1172, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] G. Charulatha et al., "Evaluation of Ground Water Quality Contaminants Using Linear Regression and Artificial Neural Network Models," *Arabian Journal of Geosciences*, vol. 10, pp. 1-9, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Song He et al., "Predictive Modeling of Groundwater Nitrate Pollution and Evaluating its Main Impact Factors Using Random Forest," *Chemosphere*, vol. 290, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Vasant Wagh et al., "Neural Network Modelling for Nitrate Concentration in Groundwater of Kadava River Basin, Nashik, Maharashtra, India," *Groundwater for Sustainable Development*, vol. 7, pp. 436-445, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Khalifa M. Alkindi et al., "Prediction of Groundwater Nitrate Concentration in a Semiarid Region Using Hybrid Bayesian Artificial Intelligence Approaches," *Environmental Science and Pollution Research*, vol. 29, pp. 20421-20436, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Raana Javidan, and Narges Javidan, "A Novel Artificial Intelligence-Based Approach for Mapping Groundwater Nitrate Pollution in the Andimeshk-Dezful Plain, Iran," *Geocarto International*, vol. 37, no. 25, pp. 10434-10458, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Asma El Amri et al., "Nitrate Concentration Analysis and Prediction in a Shallow Aquifer in Central-Eastern Tunisia Using Artificial Neural Network and Time Series Modelling," *Environmental Science and Pollution Research*, vol. 29, pp. 43300-43318, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mehdi Jamei et al., "Developing Hybrid Data-Intelligent Method Using Boruta-Random Forest Optimizer for Simulation of Nitrate Distribution Pattern," *Agricultural Water Management*, vol. 270, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Fabio Di Nunno et al., "A Nonlinear Autoregressive Exogenous (NARX) Model to Predict Nitrate Concentration in Rivers," *Environmental Science and Pollution Research*, vol. 29, pp. 40623-40642, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Hussam Eldin Elzain et al., "ANFIS-MOA Models for the Assessment of Groundwater Contamination Vulnerability in a Nitrate Contaminated Area," *Journal of Environmental Management*, vol. 286, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mosleh Hmoud Al-Adhaileh, and Fawaz Waselallah Alsaade, "Modelling and Prediction of Water Quality by Using Artificial Intelligence," *Sustainability*, vol. 13, no. 8, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Aaron Cardenas-Martinez et al., "Predictive Modelling Benchmark of Nitrate Vulnerable Zones at a Regional Scale Based on Machine Learning and Remote Sensing," *Journal of Hydrology*, vol. 603, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Raheleh Arabgol, Majid Sartaj, and Keyvan Asghari, "Predicting Nitrate Concentration and its Spatial Distribution in Groundwater Resources Using Support Vector Machines (SVMs) Model," *Environmental Modeling & Assessment*, vol. 21, pp. 71-82, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[22] AH Zare, VM Bayat, and A.P. Daneshkare, "Forecasting Nitrate Concentration in Groundwater Using Artificial Neural Network and Linear Regression Models," *International Agrophysics*, vol. 25, no. 2, pp. 187-192, 2011. [Google Scholar] [Publisher Link]

[23] Hanan Darwishe et al., "Prediction and Control of Nitrate Concentrations in Groundwater by Implementing a Model based on GIS and Artificial Neural Networks (ANN)," *Environmental Earth Sciences*, vol. 76, pp. 1-14, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[24] Farzaneh Sajedi-Hosseini et al., "A Novel Machine Learning-Based Approach for the Risk Assessment of Nitrate Groundwater Contamination," *Science of the Total Environment*, vol. 644, pp. 954-962, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] Abobakr Saeed Abobakr Yahya et al., "Water Quality Prediction Model based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios," *Water*, vol. 11, no. 6, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] M. Ehteshami, N. Dolatabadi Farahani, and S. Tavassoli, "Simulation of Nitrate Contamination in Groundwater Using Artificial Neural Networks," *Modeling Earth Systems and Environment*, vol. 2, pp. 1-10, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[27] Seiyed Mossa Hosseini, and Najmeh Mahjouri, "Developing a Fuzzy Neural Network-Based Support Vector Regression (FNN-SVR) for Regionalizing Nitrate Concentration in Groundwater," *Environmental Monitoring and Assessment*, vol. 186, pp. 3685-3699, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[28] Vasant Madhav Wagh, Dipak Baburao Panaskar, and Aniket Avinash Muley, "Estimation of Nitrate Concentration in Groundwater of Kadava River Basin-Nashik District, Maharashtra, India by Using Artificial Neural Network Model," *Modeling Earth Systems and Environment*, vol. 3, pp. 1-10, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[29] Atefeh Nouraki et al., "Prediction of Water Quality Parameters Using Machine Learning Models: A Case Study of the Karun River, Iran," *Environmental Science and Pollution Research*, vol. 28, pp. 57060-57072, 2021. [CrossRef] [Google Scholar] [Publisher Link]