Economic and Environmental Geology

THE KOREAN SOCIETY OF ECONOMIC AND ENVIRONMENTAL GEOLOGY

Research Paper

A Grey Wolf Optimized- Stacked Ensemble Approach for Nitrate Contamination Prediction in Cauvery Delta

Kalaivanan K, Vellingiri J[*]

School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore-632014, India

*Corresponding author : vellingiri.j@vit.ac.in

Research Highlights

· Nitrate contamination in river water occurs naturally and affects millions worldwide.
· Machine learning algorithms were used to predict nitrate (NO_3) contamination in river water.
· The study utilized a grey wolf optimization (GWO) algorithm to select relevant features from the dataset.
· Models were built using a stacked ensemble and four individual machine learning algorithms.
· The GWO-stacked ensemble model outperformed the others in predicting NO_3 river water contamination.

ABSTRACT

The exponential increase in nitrate pollution of river water poses an immediate threat to public health and the environment. This contamination is primarily due to various human activities, which include the overuse of nitrogenous fertilizers in agriculture and the discharge of nitrate-rich industrial effluents into rivers. As a result, the accurate prediction and identification of contaminated areas has become a crucial and challenging task for researchers. To solve these problems, this work leads to the prediction of nitrate contamination using machine learning approaches. This paper presents a novel approach known as Grey Wolf Optimizer (GWO) based on the Stacked Ensemble approach for predicting nitrate pollution in the Cauvery Delta region of Tamilnadu, India. The proposed method is evaluated using a Cauvery River dataset from the Tamilnadu Pollution Control Board. The proposed method shows excellent performance, achieving an accuracy of 93.31%, a precision of 93%, a sensitivity of 97.53%, a specificity of 94.28%, an F1-score of 95.23%, and an ROC score of 95%. These impressive results underline the demonstration of the proposed method in accurately predicting nitrate pollution in river water and ultimately help to make informed decisions to tackle these critical environmental problems.

⚒Journal homepage: http://www.kseeg.org/main.html

1. Introduction

The health of billions of people worldwide faces a significant threat due to the extensive pollution of rivers with high levels of nitrogen compounds, particularly ammonia and nitrate (NO_3) (Bagherzadeh et al., 2021). This critical issue arises from the regular consumption of river water, which often contains elevated nitrate (NO_3) concentrations, posing serious health risks. Prolonged exposure to nitrate (NO_3) in drinking water can result in a range of health conditions such as blue baby syndrome, diabetes, miscarriages, stomach cancer, and thyroid disorders (Yang et al., 2021). The detrimental impact of these health hazards is substantial, contributing to a significant portion of global diseases and cancers (Chen et al., 2017). As a result, researchers globally are actively exploring innovative approaches to address and mitigate the consequences of river water contamination (Kumar et al., 2020).

Tamil Nadu, a rapidly growing state projected to become the third most populous with over 8 million residents, faces significant water challenges. Keerthan et al. (2023) highlight that more than five million individuals in Tamil Nadu rely on the Cauvery River for their daily water requirements. However, the river water in many areas consistently exceeds the permissible nitrate (NO_3) limit of 45 mg/L (Bis, 2012) throughout the year. The Cauvery River delta, a vital agricultural region, grapples with heightened nitrate (NO_3) levels attributed to extensive nitrogen absorption from farming practices. Human-induced factors like agricultural runoff, sewage plant discharges, and nitrogenous waste oxidation in humans and animals are key contributors to the elevated nitrate (NO_3) concentrations in the Cauvery Delta region.

The regions within the Cauvery River delta exhibiting elevated nitrate (NO_3) levels also demonstrate increased concentrations of Ca, Cl, K, Mg, and Na, alongside reduced levels of SO4 (RamyaPriya et al., 2023; Tamilmani et al., 2023). Predicting nitrate (NO_3) levels accurately in river systems poses a significant challenge for environmental engineers due to the complex interplay of various factors. In response to this challenge, recent advancements in machine learning and deep learning techniques have shown promise in environmental science risk prediction. These advanced techniques excel in unravelling intricate relationships within vast datasets, handling complex patterns, and adapting continuously, offering a more robust approach compared to traditional statistical methods.

Several machine-learning methods play a crucial role in predicting river water quality, including Artificial neural networks (He et al., 2011), Adaptive network-based fuzzy inference system (Azad et al., 2018), Decision Tree (Lu et al., 2022), Random Forest (Wheeler et al., 2015), and Support vector machines (Arabgol et al., 2016). Despite the effectiveness of these techniques in water quality prediction, their application in assessing nitrate (NO_3) contamination risks remains limited, lacking an integrated approach. To address this gap, a novel framework is proposed in this study to comprehensively evaluate the risk of nitrate (NO_3) pollution. The framework focuses on developing a water quality assessment system that predicts contamination by selecting significant features to enhance classification accuracy, improve detection quality, and reduce processing time. The feature selection process relies heavily on Grey Wolf Optimization (GWO) due to its robustness and ability to identify relevant features efficiently. GWO aligns well with practical engineering challenges as it is simple, fast, precise, and easy to implement (Sharma et al., 2023). Additionally, the study introduces stacked machine learning techniques to enhance the accuracy of nitrate (NO_3)F contamination prediction, particularly when dealing with intricate datasets from diverse sources and incomplete information.

The main contributions of this study are discussed below.

- The primary aim is to introduce a novel machine-learning approach for predicting nitrate (NO_3) pollution levels in the Cauvery Delta region.
- The method proposed in this study leverages a grey wolf optimization algorithm to select relevant features from the dataset.
- Stacking, an ensemble classifier machine learning technique, is employed for task classification. This approach combines predictions from multiple base learners to enhance prediction accuracy. Each base classifier is trained to predict the reference data class, and the final model prediction is generated by the meta-learner.

• Lastly, a comparative analysis was conducted between the proposed technique and state-of-the-art methods to showcase the algorithm's effectiveness.

The rest of the paper is structured as follows: Section 2 discusses techniques for predicting water quality. Section 3 gives an introduction to GWO with Feature selection and Stacked Ensemble for water quality prediction, while Section 4 presents the results and discussion. Section 5 concludes the article.

2. Literature Review

Many research investigations have been conducted to predict nitrate contamination in rivers in India and other countries. For example, Wagh et al. (2017) proposed a technique using ANN to predict nitrate concentrations in the Kadava River catchment They collected data from 40 groundwater monitoring wells in the Nashik district and achieved an R^2 value of 0.75, indicating a good model performance. However, the dataset size is very small. Rodriguez-Galiano et al. (2018) developed a CART, RF, and SVM models to predict the relevance of characteristics associated with nitrate-related groundwater contamination. This research utilized data gathered from remote sensing technology. The Embedded, Filter, and Wrapper techniques are used to evaluate the importance of the feature. The RF-SSFS method performed better than other methods, with an AUC of 0.92. However, this study is limited to a particular area of focus. Benzer et al. (2018) created an ANN model for predicting nitrate concentrations in surface waters in a river basin in Turkey. They gathered data from 30 stations in the Yeşilırmak Watershed. The ANN model successfully predicted nitrate levels for 2020 and 2030, staying within safe drinking water standards. However, the study's limitation is that it was not tested for applicability to other regions or different contaminants. Rahmati et al. (2019) used KNN, RF, and SVM models to estimate nitrate concentration in streams in the Andimeshk-Dezful region, Iran. They used data from 114 groundwater monitoring wells in Iran and found that their RF model outperformed traditional regression models, with an R^2 of 0.72 and an RMSE of 10.41. The primary limitation of this study is based on the sampling of nitrate concentrations, assessing seasonal and interannual fluctuations in the concentrations. Knoll et al. (2019) studied different artificial intelligence methods to predict nitrate levels in groundwater in Hesse, Germany. They found that a combination of machine learning models using GBR performed best, with an R^2 of 0.75 and RMSE of 9.38 mg/l, surpassing individual models like RF, SVR, and KNN. The findings offer useful tools for water managers to forecast and control groundwater nitrate pollution, supporting environmental planning and sustainable groundwater management. However, the system they developed did not enhance accuracy. Jafari et al. (2019) created four machine learning models to forecast Total Dissolved Solids (TDS) in the Tabriz plain aquifer. These models, including ANFIS, SVM, MLP, and GEP, were trained on a dataset of 1742 groundwater samples collected from 2002 to 2012, which included various physicochemical parameters. The GEP model outperformed the others with the lowest RMSE (58.93) and the highest correlation coefficient (R = 0.998), indicating a very accurate prediction of TDS values. However, these machine learning techniques did not reduce the time complexity as expected.

Band et al. (2020) studied four machine learning models (BANN, Cubist, RF, and SVM) to predict nitrate levels in the Marvdasht watershed, Iran. They analyzed data from 67 groundwater monitoring wells and discovered that the RF model outperformed other methods with an R² of 0.89, compared to Cubist (0.87), SVM (0.74), and Bayesian-ANN (0.79). Bedi et al. (2020) compared three ML methods (ANN, XGB, and SVM) for predicting nitrate and pesticide contamination in agricultural groundwater resources. The models were assessed using a dataset consisting of 303 wells across 12 Midwestern states in the USA. The XGB model performed the best, with an RMSE value of 3.91. However, a significant limitation of this study is the scarcity of labeled data for training advanced models, which poses a challenge that requires attention in the future. Hà et al. (2020) designed an RF model to estimate nitrate and phosphorus concentrations in the Tri An reservoir. They gathered data every two months from 2009 to 2014, including parameters like TSS, TDS, COD, BOD5, EC, and turbidity. The findings demonstrated the RF model outperformed the traditional statistical methods, with an R2 value of 0.92. However, the study's limitation is that it focused solely on the specific region of the Tri An reservoir. Latif et al. (2020) developed an ANN model to forecast nitrate levels in the Feitsui reservoir in Taiwan. They input dissolved oxygen (DO), ammonium (NH3), phosphate (PO4), nitrogen

dioxide (NO2), and nitrate (NO3) parameters. The study revealed that the ANN model outperformed conventional methods, achieving an accuracy score of 0.94. However, a limitation of this study was the use of only five parameters. Stamenković et al. (2020) developed ANN and MLR models to predict the concentration of nitrates in river water. They used data from ten monitoring stations along the Danube River in Serbia from 2011 to 2016. The ANN models demonstrated good predictive ability, with a mean absolute error of 0.53 and 0.42 mg/L for the test data. However, a limitation of the study was that out of 26 parameters, only 8 showed significant deviations from the skewness and kurtosis limit values.

Alizamir et al. (2021) introduced a hybrid Bat-ELM model to forecast daily chlorophyll-a (Chl-a) levels in rivers. They used data from two USGS stations with input variables such as turbidity, pH, specific conductance, water temperature, and periodicity. The model achieved an R^2 value of 0.89. However, the key factors influencing chlorophyll-a concentration can differ based on the particular ecosystem. Pham et al. (2021) used three machine learning methods (ANN, ANFIS, and GMDH) to estimate Water Quality Index (WQI) in surface wetlands. They monitored water quality parameters like conductivity, suspended solids, BOD, ammonia, COD, dissolved oxygen, temperature, pH, phosphate, nitrite, and nitrate at seventeen wetland points over 14 months. The ANFIS method performed the best, with a low MAE of 0.0219 and a high NSE of 0.96. However, deep neural networks did not incorporate prior knowledge effectively, leading to lower prediction accuracy and longer processing times. Lu et al. (2022) developed GBRT, LSTM, and RF to predict total phosphorus and nitrogen concentrations in Taihu Lake. The data used for the study on nitrogen and phosphorus levels in Taihu Lake was gathered between 2011 and 2018, focusing on the highest monthly amounts of these substances within the lake. Results showed that the LSTM performed better compared to other models based on the RMSE value of 0.11. Ottong et al. (2022) introduced four machine learning models (LR, SVM, RF, GBM) to forecast arsenic contamination risk in the Red River Delta. They used 512 data points with 38 hadrochemical parameters from 2005 to 2007. The top-performing model was GBM, achieving high accuracy, precision, sensitivity, and specificity at 98.7%, 100%, 95.2%, and 100% respectively. One drawback of this study is the

limited amount of data. Having more data is important for creating improved models that can better adapt to different situations.

Hu et al. (2023) developed an XGB model to forecast nitrogen and phosphorus levels in Taihu lakes using 13 years of historical data. The model utilized water quality and meteorological data, achieving R2 values of 0.91 and 0.95. Sulaiman et al. (2023) compared seven machine learning models to predict nitrate concentrations for spectroscopic dataset, with the RF-PCA hybrid method performing the best at 92.7% accuracy. However, they did not choose specific features to simplify the prediction process and reduce spatial complexity. Liang et al. (2024) developed four machine learning models (GB, RF, XGB, AD) to predict nitrogen levels in Chongqing city. They analyzed 595 groundwater samples using various predictors like topography, remote sensing, hydrogeological data, climate factors, nitrate input, and socio-economic information. The GB model performed the best with an R2 of 0.627, MAE of 0.529, RMSE of 0.705, and PICP of 0.924. However, the study's limitation is that it was tested in a small area. Mehdaoui et al. (2024) introduced MLR and RBF-NN models to forecast nitrate levels in the Cheliff basin. They analyzed monthly data over a 10-year period. The RBF-NN model performed the best with an impressive accuracy of R2=0.957. One significant limitation is that the RBF-NN model is specifically suitable for this particular location. The overall literature review is summarized in Table 1, which organizes multiple articles based on selected criteria in a concise manner.

3. Methodology

The application of machine learning techniques to predict water pollution has been unsuccessful in many situations, as mentioned in the literature review section. In this study, a novel machine learning technique known as GWO-stacked ensemble learning is applied to forecast nitrate contamination, which is described below. The main objective of this work is to improve the accuracy and speed of nitrate contamination prediction by using stacked ensemble learning approaches. Stack generalization is an approach that allows researchers to combine several prediction algorithms into one. Figure 1 depicts the workflow for this study. There are various steps to the experiment. First, the Tamilnadu Pollution Control

Table 1. Literature Survey

Paper	Machine Learning Mode	Performance Metrics	Key Findings	Limitation
Wagh et al. (2017)	ANN	$R^2 = 0.75$	ANN model outperformed other methods in predicting nitrate concentrations in the Kadava River catchment	Small dataset size
Rodriguez-Galiano et al. (2018)	CART, RF, SVM	AUC = 0.92	RF-SSFS method outperformed others in nitrate-related groundwater contamination	Limited to a specific area
Benzer et al. (2018)	ANN	Accuracy = 96	ANN model effectively predicted nitrate concentrations in surface waters in a river basin in China.	The application of the model has not tested in the other regions.
Rahmati et al. (2019)	KNN, RF, SVM	$R^2 = 0.72$, RMSE = 10.41	RF model outperformed traditional regression models in estimating nitrate concentration in streams in Iran.	The model depends on the assessing seasonal and interannual fluctuations of the nitrate concentrations.
Knoll et al. (2019)	GBR, CART, MLR, RF	$R^2 = 0.75$	GBR could more accurately estimate nitrate levels	The System did not enhance accuracy
Jafari et al. (2019)	ANFIS, SVM, MLP, GEP	RMSE = 58.93, R = 0.998	GEP model provided accurate TDS prediction in Tabriz plain aquifer	Machine learning techniques did not reduce time complexity
Band et al. (2020)	BANN, Cubist, RF, SVM	$R^2 = 0.89$,	RF model outperformed others in Marvdasht watershed, Iran	Limited to a specific region
Bedi et al. (2020)	ANN, XGB, SVM	RMSE = 3.91	XGB model excelled in predicting nitrate and pesticide contamination	Scarcity of labeled data for training advanced models.
Hà et al. (2020)	RF	$R^2 = 0.92$	RF model performed better in estimating nitrate and phosphorus concentrations in Tri An reservoir	Focused solely on the Tri An reservoir.
Latif et al. (2020)	ANN	Accuracy score = 0.94	ANN model was superior in forecasting nitrate levels in Feitsui reservoir, Taiwan	Limited to five input Parameter
Stamenković et al. (2020)	ANN, MLR	MAE = 0.53	ANN models showed good predictive ability for nitrates in river water	Limited significant deviations in parameters
Alizamir et al. (2021)	Hybrid Bat-ELM	$R^2 = 0.89$	Hybrid model effectively predicted daily chlorophyll-a concentration in rivers.	Key factors influencing chlorophyll-a concentration may vary by ecosystem
Pham et al. (2021)	ANN, ANFIS, GMDH	MAE = 0.0120.0219, NSE = 0.96	ANFIS method excelled in estimating Water Quality Index in surface wetlands	Deep neural networks lacked effective incorporation of prior knowledge
Lu et al. (2022)	GBRT, LSTM, RF	RMSE = 0.11	LSTM model performed best in predicting total phosphorus and nitrogen concentrations in Taihu Lake	Focused on monthly data, limited temporal scope
Ottong et al. (2022)	LR, SVM, RF, GBM	Accuracy = 87%, Precision = 100%, Sensitivity = 95.2%, Specificity = 100%	GBM model effectively forecasted arsenic contamination risk in the Red River Delta	Limited data points for model training
Hu et al. (2023)	XGB	$R^2 = 0.91$	XGB model effectively predicted nitrogen and phosphorus concentrations in Taihu lakes.	-
Sulaiman et al. (2023)	KNN, SVM, DT, NB, RF, GB, XGB	Accuracy: 92.8%	RF-PCA hybrid method outperformed other models in predicting nitrate concentrations for hydroponic plants.	Limited Input Size
Liang et al. (2024)	GB	$R^2 = 0.627$, MAE: 0.529, RMSE: 0.705	Developed models to predict nitrogen levels in Chongqing city using various predictors	Tested in a small area
Mehdaoui et al. (2024)	RBF-NN	Accuracy = 0.957	Introduced MLR and RBF-NN models to forecast nitrate levels in the Cheliff basin	This model specifically suitable for this location
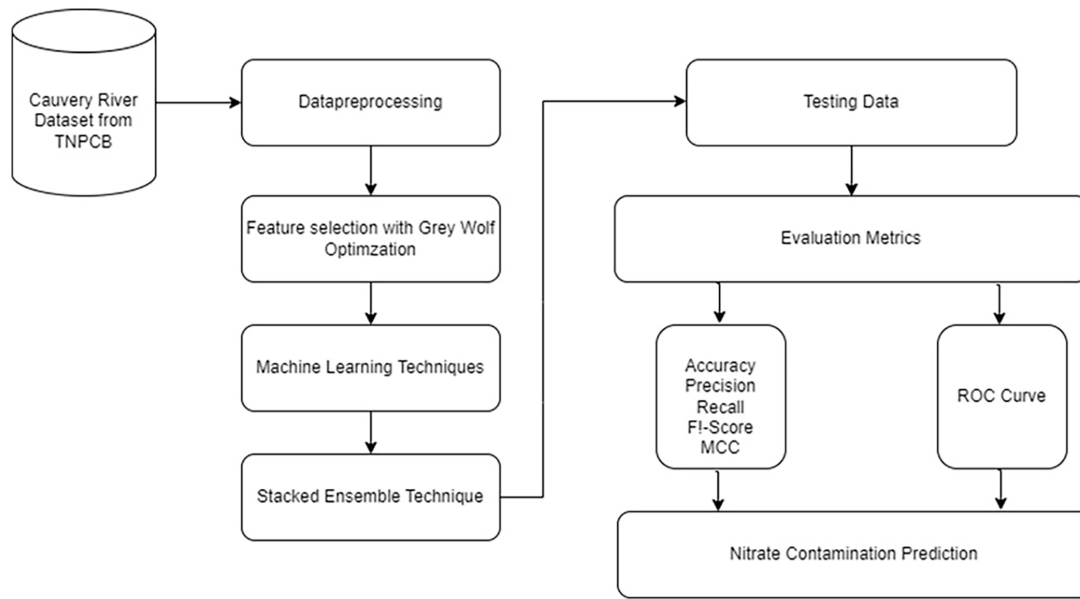
Fig. 1. System architecture of the proposed work.

Board (TNPCB) provided the dataset. It incorporates all the important water quality indicators. Water-quality data from the Cauvery River was used in this study. Typically, data pre-processing entails converting raw data into an informative format. This is a very crucial stage because datasets may contain errors, missing data, data redundancy, and noise. To solve the above issue, data pre-processing steps might be required. The next phase involves extracting relevant features via feature selection approaches using GWO. The advantages of feature selection include improving prediction accuracy, removing duplicate data from the dataset, and reducing the number of features without losing essential information. The next section compares several machine learning models, such as DT, KNN, MLP, and RF. Because each model has different classification skills, selecting the best-combined models is a difficult task in the research process. Finally, the results were assessed using several performance metrics in terms of accuracy, precision, sensitivity, specificity, F1-score, ROC, and MCC values.

3.1. Dataset Description and Preprocessing

This water quality dataset of the Cauvery River was collected by the TNPCB between 2018 and 2019. The dataset contains 792 samples and 26 features, respectively. The samples were taken at 33 monitoring sites in the Cauvery

Table 2. Attribute of water quality dataset

Variable	Description	Bureau of Indian Standard
NO_3	Nitrate	10
Ph	Potential of Hydrogen	6.5-8.5
Cl	Chloride	250
BOD	Biological oxygen demand	Not mentioned
DO	Dissolved Oxygen	Not mentioned
FC	Fecal coliforms	0.2
TC	Total coliforms	Not mentioned
Tu	Turbidity	Not mentioned
Pa	Phenolphthalein Alkalinity	Not mentioned
Tal	Total Alkalinity	200
EC	Electrical conductivity	Not mentioned
N	Nitrogen	4
COD	Chemical Oxygen Demand	Not mentioned
NH_3	Ammonia	50
Ca	Calcium	75
Th	Total hardness	300
K	Potassium	0.4
Mg	Magnesium	30
SO_4	Sulphate	200
Na	Sodium	4
TDS	Total Dissolved Solids	500
PO_4	Phosphate	Not mentioned
TFS	Total Fixed Solids	500
Br	Boron	0.3
TSS	Total Suspended Solids	500
F	Fluoride	1

River catchment area. The water quality characteristics are described in Table 2. In addition, the Z-score normalization technique is used in the data pre-processing step, which improves the quality of the dataset. Data cleaning and labeling are two steps that need to be performed before using the data. A 70–30 train-test output validation scheme was used to ensure the reliability of our test.

3.2. Grey Wolf Optimization for Feature Selection

Grey wolf optimization (GWO) was proposed by Saitali Mirzali et al. (2014) and is more successful than other optimization algorithms such as differential evolution (DE), gravity search algorithm (GSA), genetic algorithm (GA), and particle swarm optimization (PSO). GWO has been applied in many real-world applications because of its superior search ability and its use of three solutions to generate an optimal global solution (Ullah et al., 2022). This algorithm is used in a variety of applications, including wind turbines (Yang et al., 2017), feature selection (Al-Tashi et al., 2019), and image classification (Raju et al., 2018).

The algorithm is based on the social hierarchy and hunting behavior of grey wolves in the wild. The grey wolf pack has a rigid social structure comprising alpha (α), beta (β), delta (δ), and omega (Figure 2). As pack leader, the alpha wolf assigns tasks to the other wolves. The beta wolf acts as a bridge between the alpha wolf and the other wolves in the pack, and its position can help the other wolves explore new regions in the search space. Delta wolves are called the heart of the pack, and their main job is hunting. The Omega wolves are at the bottom of the swarm and mostly serve as babysitters. Figure 3 is a flowchart explaining the operation of GWO.
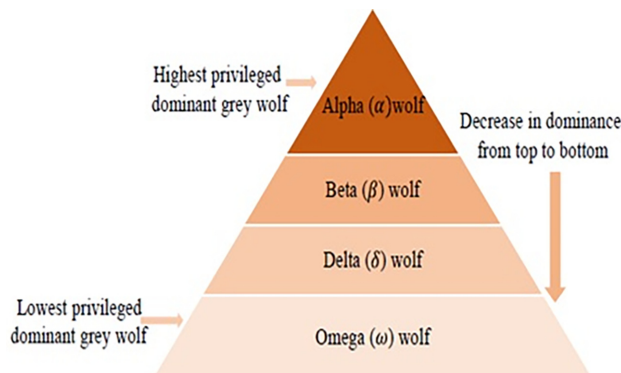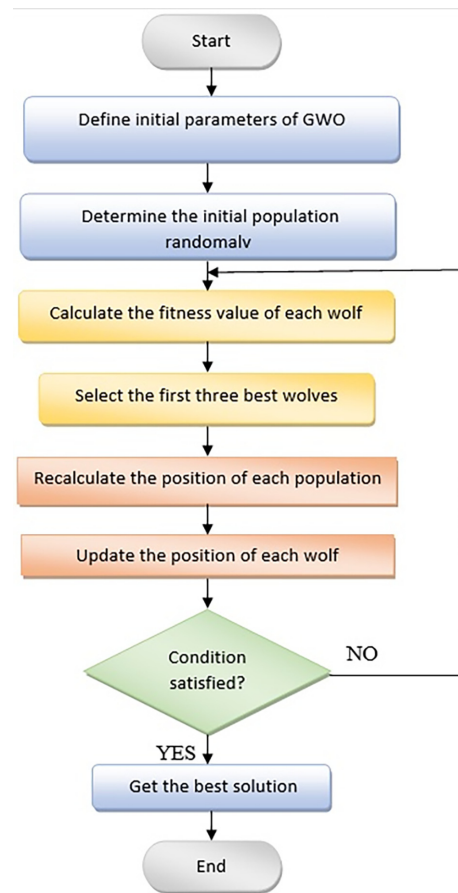

Fig. 2. Hierarchy of wolves.

Fig. 3. Flowchart for working principle of GWO.

The Grey wolf position vector may be defined as

$$\vec{W} = \{W_1, W_2.....W_n\} \tag{1}$$

In GWO, the hunting process behavior is described as follows

$$\vec{P} = \left| \vec{B} * \vec{W}_p(z) - \vec{W}(z) \right| \tag{2}$$

$$\vec{W}(z+1) = \vec{W}_p(z) - \vec{A} * \vec{X} \tag{3}$$

Where z = current iteration, $\vec{W}_p(z)$ = the prey position, $\vec{W}$ = grey wolf position vector. The parameters $\vec{A}, \vec{B}$ are computed as follows

$$\vec{A} = 2\vec{a}\vec{s}_1 - \vec{a}$$
$$\vec{B} = 2\vec{s}_2 \tag{4}$$

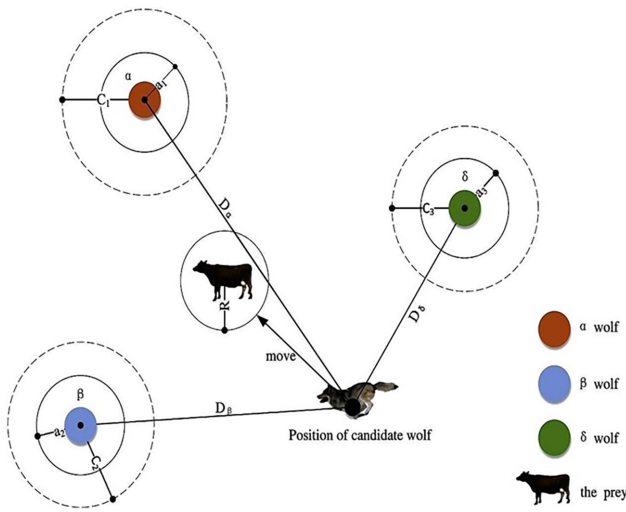Where s1 and s2 are randomly initialized variables and

Fig. 4. Update position in GWO during the Hunting Process.

represent a decrease in iteration from 2 to 0.

The presence of alpha, beta, and delta wolves in the hunting area has caused the status of grey wolves to be adjusted according to their relative positions to these wolves. Figure 4 illustrates the updated status of grey wolves in the hunting section.

$$\vec{P}_{\alpha} = \left| \vec{B}_1 * \vec{W}_{\alpha} - \vec{W} \right|$$

$$\vec{P}_{\beta} = \left| \vec{B}_2 * \vec{W}_{\beta} - \vec{W} \right|$$

$$\vec{P}_{\delta} = \left| \vec{B}_3 * \vec{W}_{\delta} - \vec{W} \right| \tag{5}$$

$$\vec{W}_1 = \vec{W}_{\alpha} + A_1 * \vec{P}_{\alpha}$$

$$\vec{W}_2 = \vec{W}_{\beta} + A_2 * \vec{P}_{\beta}$$

$$\vec{W}_3 = \vec{W}_{\delta} + A_3 * \vec{P}_{\delta} \tag{6}$$

$$W(z+1) = \frac{\vec{W}_1 + \vec{W}_2 + \vec{W}_3}{3} \tag{7}$$

where W_{α}= the position of alpha wolves. W_{β}= the position of beta wolves, W_{δ}= the position of delta wolves.

3.3. Classifiers

The machine learning techniques DT, MLP, KNN, RF, and Stacked ensemble were used to predict water quality to accomplish this objective.

Decision Tree (DT): -The DT has three distinct components - an inner node, a branch node, and a leaf node - that function similarly to a traditional tree. Each inner node acts as a test variable, each branch indicates the result of the test, and each leaf node contains the class label. The entropy technique is employed to select the variable that will serve as the root of a decision tree. The tree is then divided into multiple subsets based on the values of the test attributes. This recursive approach is performed for each subset until they are all resolved. This recursive partitioning procedure separates the population into subpopulations depending on dichotomous variables, yielding a decision tree that appropriately identifies each person. (Myles et al., 2004).

KNN Algorithms: -KNN is a sluggish machine learning technique that can be utilized for classification and regression problems. This algorithm is widely used in data mining, pattern recognition, and intrusion detection. This approach uses distance calculations to provide unique predictions based on data that has been observed. The most commonly used methods for this calculation are the Euclidean distance, the Mahalanobis distance, and the city-block distance. The K number of points is usually determined by how close the test data is to the known points. The advantage of KNN classification is its simplicity and non-parameter.

Multi-layer perceptron (MLP): -MLP is a kind of feed-forward ANN comprising a single-layer perceptron. An entry layer, a hidden layer, and an exit layer are three components used to create MLP. MLP has been used as a front propagation learning technique to transmit data from an input node to an output node. The learning capacity of MLP is determined by connection weights. The performance of the network increases over time by repeatedly adjusting the connection weights (Atangana et al., 2020; Joy et al., 2020). MLP is a supervised ML technique that is mostly used to classify patterns (Guo et al., 2020).

Random Forest: -RF is a supervised type of ML technique for regression and classification. It comprises several decision trees that depend on either the bagging or bootstrap aggregating approach. Random forest is used in ensemble learning techniques to solve complex problems

and increase accuracy by merging individual models. The overall vote of all trees determines the final classification outcome (Chen et al., 2020).

3.4. Stacked-generalization Model

Stacked is an ensemble learning approach that involves the integration of multiple base models to improve the overall prediction of machine learning. It is a higher-level approach to combine models compared to techniques, such as bagging and boosting, which focus on creating multiple models with different random subsets of the data or modifying the weights of training examples. The basic idea behind stacking is to use a set of diverse base models that are trained on different subsets of the data, using different algorithms and hyperparameters. Each base model makes its predictions, which are then combined with the meta-model to produce a final output. Figure 5 depicts the general form of the proposed stacked ensemble model. Random forest, Multiple-layer perception. Decision tree and KNN are the models used in the research study.

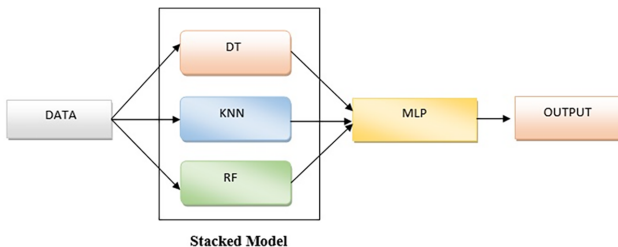The pseudo-code for the stacked ensemble technique is given below

Dataset (D) =$\{(x_1,y_1),(x_2,y_2)\ldots\ldots\ldots(x_z,y_z)$
z: the size of the dataset
n: number of base learners
Level_0 or Base Learners(b): {DT, KNN, RF}
Level_1 or Meta Learners(m): {MLP}

Procedure:
Step1: Learn base classifiers based on D
    For (i=1 to z)
        For (n=1 to 3)
            $P_{in}=b_n (x_i)$
    End For

Stacked Model

Fig. 5. The stacked model with meta learner = Multiple-layer Perception and the weak learners = Decision tree, Random Forest, and K-nearest neighbor.

Step 2: Constructing a new dataset based on predictions
Generate an empty dataset (D_1)
For (i=1 to Z)
$D_1 = D_1 \ U\{((p_{i1},p_{i2},\ldots p_{13}),yi)\}$
End For

Step 3: Learn Level_1 Learners Based on D_1
B1= M (D1)

Output
$B_1(X)= B_1 (b_1(x),b_2(x),b_3(x))$
$b_n(D)$
If (n= =1): Apply DT on D
Else If (n= =2): Apply KNN on D
Else If (n= =3): Apply RF on D
M (D_1)
For (i=1 to Z)

3.5. NO$_3$ Prediction Procedure

The following technique was performed for a hybrid GWO-stacked ensemble

Step 1: Collect the Cauvery River data from the Tamilnadu Pollution Control Board.
Step 2: Data pre-processing techniques are implemented using Z score normalization.
Step 3: The GWO feature selection approach is used to extract the essential features from the dataset.
Step 4: Divide the dataset into train and test sets.
Step 5: Training samples are analyzed using the stacked ensemble classification algorithm.
Step 6: The trained classifier is used on experimental data samples to predict whether NO3 contamination is at an acceptable level or not.
Step 7: Finally, the results recommend a suitable model for the prediction of NO3.

4. Result and Discussion

4.1. Model Evaluation and Experimental Setup

All experiments in this study were conducted with Python using the Jupyter Notebook framework on a Dell laptop, Intel Core™ i5-10210U CPU @ 1.60 GHz and 16 GB RAM. Pandas, NumPy, and Matplotlib libraries were used. The performance of the GWO-stacked model was evaluated

using the following metrics, represented mathematically as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (8)$$

$$Precision = \frac{TP}{TP + FP} \qquad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad (10)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (11)$$

$$F1\text{-}SCORE = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (12)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (13)$$

4.2. Selection of Input Features

Feature selection is a crucial step in the machine learning pipeline. It involves selecting a subset of relevant features from the original dataset to improve the performance and interpretability of the model. To show the superiority of the proposed GWO stacked ensemble method, four standalone ML models were also tested and used to compare their performances with those of the GWO stacked ensemble. To perform sensitivity analyses faster, the results were experimentally performed considering different input variables,

i.e. BOD, Ca, Cl, K, Mg, Na, NH3, N, and S04 the best prediction accuracies were obtained with the GWO stacked ensemble.

4.3. Evaluation Process

The experimental results of the GWO stacked ensemble method are evaluated in comparison with different machine learning techniques such as DT, KNN, MLP, and RF. The GWO stacked ensemble method is tested in the Python environment using the Cauvery River data obtained from TNPCB. The results of the confusion matrix of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are shown in Table 3. This value was used to determine the performance of a classification on test data. The GWO stacked ensemble method had the highest TP and the lowest FN of 76 and 6, respectively. In addition, the negative results were also perfectly predicted with a TN and FP of 76 and 4, respectively. This indicates that the GWO stacked ensemble technique has the best hyperparameters compared to other methods.

Table 4 shows that our proposed model performs well compared to all other models in terms of other performance parameters. For the Matthew coefficient, the RF classifier achieved the second-highest score of 80%. The findings demonstrated no significant differences between DT and MLP, with precision and specificity ratings of 86% and 83%, respectively. However, the KNN classifier has the lowest accuracy, sensitivity, F1-score, and ROC scores, with

Table 3. Confusion matrix result for test data

S.NO	Classifier	TP	FP	FN	TN	Accuracy
1	GWO-Stacked Ensemble (Proposed)	76	4	6	75	0.93
2	RF	73	6	7	76	0.90
3	DT	71	8	9	73	0.88
4	MLP	72	10	10	69	0.87
5	KNN	70	10	11	69	0.86

Table 4. Comparison of the proposed model with the base classifiers

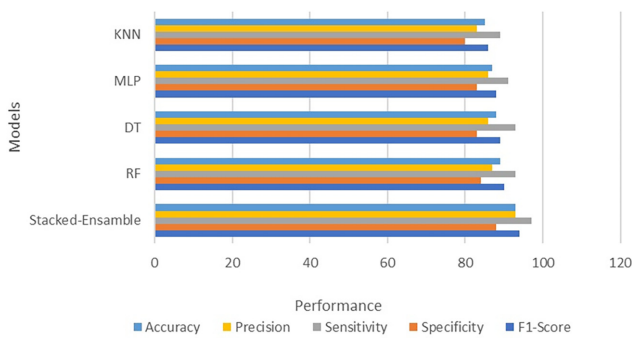| Classifier	Accuracy	Precision	Sensitivity	Specificity	F1-Score	MCC
GWO-Stacked Ensemble (Proposed)	0.93	0.95	0.92	0.92	0.93	0.85
RF	0.90	0.92	0.91	0.90	0.91	0.83
DT	0.88	0.89	0.88	0.89	0.89	0.78
MLP	0.87	0.87	0.86	0.85	0.87	0.74
KNN	0.86	0.87	0.87	0.86	0.86	0.72

Fig. 6. Comparison graph of the proposed approach with different base classifier on the Cauvery dataset.

values of 85%, 89%, 86%, and 90%, respectively. Figure 6 shows the performance comparison of each model. In this figure, the GWO-stacked model outperforms other ML models in many situations. Figure 3 shows the ROC curve of the predictive performance of all models. According to the graph, the GWO-stacked reached the maximum value of 0.95.

## 4.4. Performance as of the Proposed Method with the Base Classifier Using Data Splitting Validation

Table 5 presents a performance assessment of the proposed method with base classifier methods to select the best

Table 5. Comparison of the performance of the proposed methods against the basic classifier using data splitting validation

Metrics	Data Split Ratio	Classification Method				
		GWO-Stacked (Proposed)	RF	DT	MLP	KNN
Accuracy	60-40	91.5	88	86.5	85.3	83.55
	70-30	93.21	89.78	88	87.65	85.52
	80-20	92.98	89.16	87.05	87.23	84.89
	90-10	91.94	88.94	86.83	85.32	82.94
Precision	60-40	91.56	85.09	84.78	84.35	81.36
	70-30	93	87.26	86.72	86.72	83.88
	80-20	92.13	86.58	85.09	85.39	81.36
	90-10	91.04	85.98	84.21	83.96	80
Sensitivity	60-40	95.13	91.89	91.27	89.28	87.18
	70-30	97.53	93.33	93	91.86	89.10
	80-20	96.41	91.56	91.04	89.74	88.72
	90-10	95.23	90.86	90.12	88.21	87.23
Specificity	60-40	86.94	82.92	81.89	81.11	78.27
	70-30	88.56	84.23	83.98	83.15	80
	80-20	87.88	83.55	82.45	82.18	78.89
	90-10	86.52	82.10	81.89	81.23	77.25
F1-Score	60-40	92.41	88.72	86.94	85.36	85.10
	70-30	94.28	90	89.56	88.78	86.23
	80-20	93.88	88.72	87.90	87.45	85.63
	90-10	92.58	86.28	85.92	85.11	84.23
ROC	60-40	93.88	92.12	90.25	89.41	88.23
	70-30	95.23	94.15	92.10	91	90.05
	80-20	94.73	93.78	91.65	89.41	88.14
	90-10	93.12	92.89	92.10	88.11	87.23
MCC	60-40	81.56	77.89	76.23	75.96	72.18
	70-30	83	79.52	78.36	75.12	74.89
	80-20	82.16	78.23	77.65	73.96	73.98
	90-10	81.23	77.23	76.63	72.13	71.08

Table 6. Shows a comparison of performance between the Proposed method and existing research

Author	Model	Accuracy (%)
Proposed Model	GWO-Stacked Ensemble	93
Latif et al. (2020)	ANN	93
Sulaiman et al. (2023)	KNN, SVM, DT, NB, RF, GB, XGB	92.8
Bhattarai et al. (2021)	KNN, NB, RF, GB, SVM	92.8
Alizamir et al. (2021)	Hybrid Bat-ELM	89
Knoll et al. (2019)	GBR, CART, MLR, RF	75
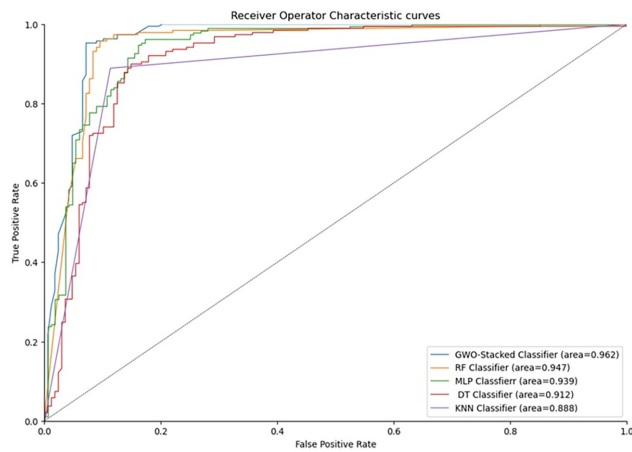
Fig. 7. ROC curve of the Stacked Ensemble in comparison to the basic classifier on the Cauvery River dataset.

model for predicting nitrate contamination. The optimal model is determined by dividing the data into test and training ranges. The ranges vary from 60%–40% to 90%–10%. The performance is assessed using different evaluation metrics. Table 5 demonstrates the performance of the GWO-stacked algorithm compared to other studies in this field when the data is split at a ratio of 70:30. The outcomes of the GWO-stacked method are assessed using the seven metrics listed above, and this information is used to determine the best data separation threshold for predicting nitrate pollution in river water.

4.5. Performance as of the Proposed Method with the Other Classifier Techniques

Table 6 compares the accuracy of various advanced techniques with the proposed system, showing that GWO-Stacked had the highest accuracy in predicting nitrate concentration. Latif et al. (2020) achieved the highest accuracy of 93% using ANN, while Sulaiman et al. (2020) and Bhattarai et al. (2021) reached 92.8% accuracy each.

On the other hand, Alizamir et al. (2021) and Knoll et al. (2019) had models with less than 90% accuracy in predicting nitrate concentration.

5. Conclusion

In this study, a machine learning approach called the GWO-stacked ensemble is proposed for predicting nitrogen pollution in the Cauvery River Delta region. The model involves data preprocessing to handle missing values and normalization, followed by feature selection using the grey wolf optimization technique. This method efficiently selects relevant features for input into the stacked ensemble algorithm, which mitigates issues like variance and overfitting seen in single-classifier models. The GWO-stacked ensemble outperformed DT, RF, MLP, and KNN models with an accuracy of 93%, precision of 93%, sensitivity of 97%, specificity of 88%, and F1-score of 94%. The ROC curve accuracy was highest at 95% with this technique. The research though it achieved its goals is limited by its reliance on a few factors. This narrow focus helps forecast levels even when data is scarce enhancing the usefulness of the models. Therefore, it's important for future studies to identify factors that could enhance the power of machine learning algorithms in this specific field.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, [Vellingiri. J], upon reasonable request.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Reference

Alizamir, M., Heddam, S., Kim, S. and Mehr, A.D. (2021). On the implementation of a novel data-intelligence model based on extreme learning machine optimized by bat algorithm for estimating daily chlorophyll-a concentration: case studies of river and lake in USA. J. Clean Prod., v.285, 124868. doi: 10.1016/j.jclepro.2020.124868

Al-Tashi, Q., Kadir, S. J. A., Rais, H. M., Mirjalili, S. and Alhussian, H. (2019). Binary optimization using hybrid grey wolf optimization for feature selection. IEEE Access, v.7, p.39496-39508. doi: 10.1109/ACCESS.2019.2906757

Arabgol, R., Sartaj, M. and Asghari, K. (2016). Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model. Environmental Modeling & Assessment, v.21, p.71-82. doi: 10.1007/s10666-015-9468-0

Atangana, R., Tchiotsop, D., Kenne, G. and Chanel, L. (2020). EEG signal classification using LDA and MLP classifier. Health Informat. Int. J., v.9(1), p.14-32. doi: 10.5121/hiij.2020.9102

Azad, A., Karami, H., Farzin, S., Saeedian, A., Kashi, H. and Sayyahi, F. (2018). Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River). KSCE Journal of Civil Engineering, v.22(7), p.2206-2213 doi: 10.1007/s12205-017-1703-6

Bagherzadeh, F., Mehrani, M.J., Basirifard, M. and Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. Journal of Water Process Engineering, v.41, 102033. doi: 10.1016/j.jwpe.2021.102033

Band, S.S., Janizadeh, S., Pal, S.C., Chowdhuri, I., Siabi, Z., Norouzi, A., ... and Mosavi, A. (2020). Comparative analysis of artificial intelligence models for accurate estimation of groundwater nitrate concentration. Sensors, v.20(20), 5763. doi: 10.3390/s20205763

Bedi, S., Samal, A., Ray, C. and Snow, D. (2020) Comparative evaluation of machine learning models for groundwater quality assessment. Environ Monit Assess, v.192(12), p.1-23. doi: 10.1007/s10661-020-08695-3

Benzer, S. and Benzer, R. (2018). Modelling nitrate prediction of groundwater and surface water using artificial neural networks. Politeknik Dergisi, v.21(2), p.321-325. doi: 10.2339/politeknik.385434

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

Bis, I. (2012). 10500 Indian standard drinking water–specification, second revision. Bureau of Indian Standards, New Delhi.

Chen, J., Wu, H., Qian, H. and Gao, Y. (2017). Assessing nitrate and fluoride contaminants in drinking water and their health risk of rural residents living in a semiarid region of Northwest China. Exposure and Health, v.9(3), p.183-195. doi: 10.1007/s12403-016-0231-9

Dokala, J.K.K., Mohamed, M.R., Kumarasamy, S. and Kurukuri, P. (2022). A new meta-heuristic optimization algorithm based MPPT control technique for PV System under diverse partial shading conditions. doi: 10.21203/rs.3.rs-1531369/v1

Guo, L., Rivero, D., Dorado, J., Rabunal, J.R. and Pazos, A. (2010). Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks. Journal of Neuroscience Methods, v.191(1), p.101-109. doi: 10.1016/j.jneumeth.2010.05.020

Ha, N.T., Nguyen, H.Q., Truong, N.C.Q., Le, T.L., Thai, V.N. and Pham, T.L. (2020). Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri An Reservoir, Vietnam. Environmental Monitoring and Assessment, v.192, p.1-20. doi: 10.1007/s10661-020-08731-2

He, B., Oki, T., Sun, F., Komori, D., Kanae, S., Wang, Y., ... and Yamazaki, D. (2011). Estimating monthly total nitrogen concentration in streams by using artificial neural network. Journal of Environmental Management, v.92(1), p.172-177. doi: 10.1016/j.jenvman.2010.09.014

Hu, Y., Du, W., Yang, C., Wang, Y., Huang, T., Xu, X. and Li, W. (2023). Source identification and prediction of nitrogen and phosphorus pollution of Lake Taihu by an ensemble machine learning technique. Frontiers of Environmental Science & Engineering, v.17(5), 55. doi: 10.1007/s11783-023-1655-7

Jafari, R., Torabian, A., Ghorbani, M.A., Mirbagheri, S.A. and Hassani, A.H. (2019) Prediction of groundwater quality parameter in the Tabriz plain, Iran using soft computing methods. J. Water Suppl. Res. Technol.— AQUA, v.68(7), p.573-584. doi: 10.2166/aqua.2019.062

Joy, J., Kannan, A., Ram, S. and Rama, S. (2020). Speech emotion recognition using neural network and MLP classifier. IJESC, April-2020.

Keerthan, L., RamyaPriya, R. and Elango, L. (2023). Geogenic and anthropogenic contamination in river water and groundwater of the lower Cauvery Basin, India. Frontiers in Environmental Science, v.11, 278. doi: 10.3389/fenvs.2023.1001052

Knoll, L., Breuer, L. and Bach, M. (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of the Total Environment, v.668, p.1317-1327. doi: 10.1016/j.scitotenv.2019.03.045

Kumar, P., Lai, S.H., Mohd, N.S., Kamal, M.R., Afan, H.A., Ahmed, A.N., ... and El-Shafie, A. (2020). Optimised neural network model for river-nitrogen prediction utilizing a new training approach. PLoS One, v.15(9), e0239509. doi: 10.1371/journal.pone.0239509

Latif, S.D., Azmi, M.S.B.N., Ahmed, A.N., Fai, C.M. and El-Shafie, A. (2020). Application of artificial neural network for forecasting nitrate concentration as a water quality parameter: a case study of Feitsui Reservoir, Taiwan. Int. J. Des. Nat. Ecodynamics, v.15, p.647-652. doi: 10.18280/ijdne.150505

Liang, Y., Zhang, X., Gan, L., Chen, S., Zhao, S., Ding, J., ... and Yang, H. (2024). Mapping specific groundwater nitrate concentrations from spatial data using machine learning: A case study of chongqing, China. Heliyon. doi: 10.1016/j.heliyon.2024.e27867

Lu, H., Yang, L., Fan, Y., Qian, X. and Liu, T. (2022). Novel simulation of aqueous total nitrogen and phosphorus concentrations in Taihu Lake with machine learning. Environmental Research, v.204, 111940. doi: 10.1016/j.envres.2021.111940

Lu, H., Yang, L., Fan, Y., Qian, X. and Liu, T. (2022). Novel simulation of aqueous total nitrogen and phosphorus concentrations in Taihu Lake with machine learning. Environmental Research, v.204, 111940. doi: 10.1016/j.envres.2021.111940

Mehdaoui, I., Boudibi, S., Latif, S.D., Sakaa, B., Chaffai, H. and Hani, A. (2024). Prediction of nitrate concentrations using multiple linear regression and radial basis function neural network in the Cheliff River basin, Algeria. Journal of Applied Water Engineering and Research, v.12(1), p.77-89. doi: 10.1080/23249676.2023.2207838

Mirjalili, S., Mirjalili, S.M. and Lewis, A. (2014). Grey wolf optimizer. J. Advances in Engineering Software. v.69, p.46-61. doi: 10.1016/j.advengsoft.2013.12.007

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. and Brown, S.D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, v.18(6), p.275-285. doi: 10.1002/cem.873

Ottong, Z.J., Puspasari, R.L., Yoon, D. and Kim, K.W. (2022). Predicting As Contamination Risk in Red River Delta using Machine Learning Algorithms. Available at SSRN 3952430. doi: 10.9719/EEG.2022.55.2.127

Pham, Q.B., Mohammadpour, R., Linh, N.T.T., Mohajane, M., Pourjasem, A., Sammen, S.S. and Anh, D.T. (2021). Application of soft computing to predict water quality in wetland. Environ. Sci. Pollut. Res., v.28(1), p.185-200. doi: 10.1007/s11356-020-10344-8

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., ... and Bui, D.T. (2019). Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Science of the Total Environment, v.688, p.855-866. doi: 10.1016/j.scitotenv.2019.06.320

Raju, P., Rao, V.M. and Rao, B.P. (2018). Grey wolf optimization-based artificial neural network for classification of kidney images. Journal of Circuits, Systems and Computers, v.27(14), 1850231. doi: 10.1142/S0218126618502316

RamyaPriya, R. and Elango, L. (2021). Atmospheric CO_2 consumption by rock weathering over a five year period in a large non-perennial tropical river basin of southern India. Environmental Science and Pollution Research, v.28, p.26461-26478. doi: 10.1007/s11356-020-12257-y

Rodriguez-Galiano, V.F., Luque-Espinar, J.A., Chica-Olmo, M. and Mendes, M.P. (2018). Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Science of the Total Environment, v.624, p.661-672. doi: 10.1016/j.scitotenv.2017.12.152

Sharma, S. and Yadav, N.S. (2023). A Grey Wolf Optimized XGboost-Multilayer Stacking Approach to Detect XSS Attacks. IETE Journal of Research, 1-16. doi: 10.1080/03772063.2023.2264251

Stamenković, L.J., Mrazovac Kurilić, S. and Presburger Ulniković, V. (2020). Prediction of nitrate concentration in Danube River water by using artificial neural networks. Water Supply, v.20(6), p.2119-2132. doi: 10.2166/ws.2020.104

Sulaiman, R., Azeman, N.H., Bakar, M.H.A., Nazri, N.A.A., Masran, A.S. and Bakar, A.A.A. (2023). Nitrate Classification Based on Optical Absorbance Data Using Machine Learning Algorithms for a Hydroponics System. Applied Spectroscopy, v.77(2), p.210-219. doi: 10.1177/00037028221140924

Tamilmani, A. and Venkatesan, G. (2021). Assessment of trace metals and its pollution load indicators in water and sediments between Upper and Grand Anicuts in the Cauvery. International Journal of Environmental Science and Technology, 1-12. doi: 10.1007/s13762-020-03034-y

Ullah, I., Liu, K., Yamamoto, T., Shafiullah, M. and Jamal, A. (2022). Grey wolf optimizer-based machine learning algorithm to predict electric vehicle charging duration time. Transportation Letters, 1-18. doi: 10.1080/19427867.2022.2111902

Wagh, Vasant Madhav, Dipak Baburao Panaskar, and Aniket Avinash Muley. (2017). 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, v.3, p.1-10. doi: 10.1007/s40808-017-0290-3

Wheeler, D.C., Nolan, B.T., Flory, A.R., DellaValle, C.T. and Ward, M.H. (2015). Modeling groundwater nitrate concentrations in private wells in Iowa. Science of the Total Environment, v.536, p.481-488. doi: 10.1016/j.scitotenv.2015.07.080

Yang, B., Zhang, X., Yu, T., Shu, H. and Fang, Z. (2017). Grouped grey wolf optimizer for maximum power point tracking of doubly-fed induction generator-based wind turbine. Energy Conversion and Management, v.133, p.427-443. doi: 10.1016/j.enconman.2016.10.062

Yang, Y., Shang, X., Chen, Z., Mei, K., Wang, Z., Dahlgren, R.A., ... and Ji, X. (2021). A support vector regression model to predict nitrate-nitrogen isotopic composition using hydro-chemical variables. Journal of Environmental Management, v.290, 112674. doi: 10.1016/j.jenvman.2021.112674