



Hydrogeochemistry and prediction of arsenic contamination in groundwater of Vehari, Pakistan: comparison of artificial neural network, random forest and logistic regression models

Javed Iqbal · Chunli Su · Maqsood Ahmad · Muhammad Yousuf Jat Baloch ·
Abdur Rashid · Zahid Ullah · Hasnain Abbas · Anam Nigar · Asmat Ali ·
Arif Ullah

Received: 2 June 2023 / Accepted: 10 October 2023 / Published online: 26 December 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract Arsenic contamination in the groundwater occurs in various parts of the world due to anthropogenic and natural sources, adversely affecting human health and ecosystems. The current study intends to examine the groundwater hydrogeochemistry containing elevated arsenic (As), predict As levels in groundwater, and determine the aptness of groundwater for drinking in the Vehari district, Pakistan. Four hundred groundwater samples from the study region were collected for physiochemical analysis. As levels in groundwater samples ranged from 0.1 to 52 µg/L, with an average of 11.64 µg/L, (43.5%), groundwater samples exceeded the WHO 2022 recommended limit of 10 µg/L for drinking purposes.

Ion-exchange processes and the adsorption of ions significantly impacted the concentration of As. The HCO_3^- and Na^+ are the dominant ions in the study area, and the water types of samples were CaHCO_3 , mixed CaMgCl , and CaCl , demonstrating that rock-water contact significantly impacts hydrochemical behavior. The geochemical modeling indicated negative saturation indices with calcium carbonate and other salt minerals, encompassing aragonite, calcite, dolomite, and halite. The dissolution mechanism suggested that these minerals might have implications for the mobilization of As in groundwater. A combination of human-induced and natural sources of contamination was unveiled through principal component analysis (PCA). Artificial neural networks (ANN), random forest (RF), and logistic regression (LR) were used to predict As in the groundwater.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10653-023-01782-7>.

J. Iqbal · C. Su (✉) · A. Rashid · Z. Ullah · H. Abbas ·
A. Ali
School of Environmental Studies, China
University of Geosciences, Wuhan 430074,
People's Republic of China
e-mail: chl.su@cug.edu.cn

J. Iqbal · C. Su · A. Rashid · Z. Ullah · H. Abbas · A. Ali
State Environmental Protection Key Laboratory of Source
Apportionment and Control of Aquatic Pollution, China
University of Geosciences, Wuhan 430074, China

M. Ahmad
School of Geography and Information Engineering, China
University of Geosciences, Wuhan 430074, China

M. Y. J. Baloch
College of New Energy and Environment, Jilin University,
Changchun 130021, China

A. Nigar
School of Electronics and Information Engineering,
Changchun University of Science and Technology,
Changchun 130022, China

A. Ullah
Institute of Geological Survey, China University
of Geosciences, 388 Lumo Road, Wuhan 430074, China

The data have been divided into two parts for statistical analysis: 20% for testing and 80% for training. The most significant input variables for As prediction was determined using Chi-squared analysis. The receiver operating characteristic area under the curve and confusion matrix were used to evaluate the models; the RF, ANN, and LR accuracies were 0.89, 0.85, and 0.76. The permutation feature and mean decrease in impurity determine ten parameters that influence groundwater arsenic in the study region, including F^- , Fe^{2+} , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , SO_4^{2-} , NO_3^- , HCO_3^- , and Na^+ . The present study shows RF is the best model for predicting groundwater As contamination in the research area. The water quality index showed that 161 samples represent poor water, and 121 samples are unsuitable for drinking. Establishing effective strategies and regulatory measures is imperative in Vehari to ensure the sustainability of groundwater resources.

Keywords Groundwater arsenic · Machine learning · Logistic regression · Artificial neural network · Random forest · Water quality index

Introduction

People rely heavily on groundwater in many countries for drinking, agriculture, and industrial needs (Jat Baloch et al., 2020; Rehman et al., 2019; Ullah et al., 2022c). Groundwater supplies drinking water for one-third of the world and is the source of freshwater in arid and semi-arid areas of Pakistan (Ghani et al., 2022; Jat Baloch et al., 2023). Because of rapid population expansion, agricultural and industrial activities, groundwater withdrawal has steadily increased and prompted concerns regarding assessing and managing groundwater resources for sustainable development (Iqbal et al., 2023b; Rashid et al., 2023; Ullah et al., 2022a). Groundwater is one of the essential water resources in Pakistan (Jat Baloch et al., 2020, 2023). However, several chemical elements, such as arsenic (As), threaten groundwater quality increasingly. Over 47 million are currently exposed to As contamination in Pakistan (Jat Baloch et al., 2022a; Rashid et al., 2023). Thus, understanding groundwater quality is critical for effective water management and long-term sustainability (Zhang et al., 2022; Zhang et al., 2022b).

The International Agency for Research on Cancer (IARC) classifies As as a class 1 human toxic element (Tropea et al., 2021). The World Health Organization (WHO) has reduced the As level in drinkable water from 50 to 10 $\mu\text{g/L}$ due to the high carcinogenic risk (Zhou et al., 2021). Water contamination is a significant concern worldwide, notably in developing countries such as China, Pakistan, India, Bangladesh, and Vietnam. The consumption of As in drinking water has affected over 2 million people worldwide (Rahman et al., 2021). Weathering, evapotranspiration, and volcanic emissions are all geological factors that influence groundwater quality. Other recent studies have found that human activities contribute significantly to groundwater contamination. Subsurface contamination is caused by activities such as petroleum refining, herbicides, pesticides, and mining (Dilpazeer et al., 2023; Li et al., 2023; Stojanović Bjelić et al., 2023; Tariq et al., 2023). As contamination in eggs, water, milk, food, and meat can result in many health problems. The ingestion of bovine milk is among the most significant sources of toxicants in the food chain (Ullah et al., 2021). Groundwater As concentration rises due to physicochemical and geochemical conditions and rock-water interaction. People can be exposed to As through various mechanisms, including breathing, drinking, and skin contact (Çiner et al., 2021). Numerous national and international cases demonstrate that drinking contaminated water endangers people's health via these interconnected pathways (Iqbal et al., 2023a; Tabassum et al., 2019). This high As concentration in drinking water may cause various health issues, including hair loss, kidney failure, and cardiovascular disease (Rashid et al., 2019). Geochemical compositions, concentration levels, and bedrock geology all have an impact on groundwater quality around the world. Freshwater resources are critical for all life forms and are required for the survival of life and the natural environment. Overconsumption and poor management threaten freshwater resources (Jat Baloch & Mangi, 2019). To identify trends and ensure sustainability, groundwater modeling, quality analysis, and monitoring are required.

In recent water studies, Machine Learning (ML) methods are often used to solve various issues (Hussain et al., 2022; Sahin et al., 2021; Sun & Scanlon, 2019). ML approaches generally emphasize the relationship between the model's outputs and inputs

rather than the mechanisms that enable the process. Sophisticated nonlinear associations between many variables can be appropriately documented with or without previous knowledge of the investigated system by learning a massive amount of data (Abbas et al., 2023; Hussain et al., 2021; Iqbal et al., 2020; Jamil et al., 2019). The presence of F^- , As, and other contaminants in groundwater has thus been estimated using various ML techniques, such as Artificial neural networks (ANN) (Ahmadi et al., 2017). The Random Forest (RF) model is most widely used for regression and classification. RF has many valuable features for classification. Because RF is a non-parametric, nonlinear method, it can handle large datasets with numerical and categorical data and complicated non-linearity and factor interactions (Ranjgar et al., 2021).

Furthermore, logistic regression defines and clarifies the relationship between one or more independent nominal, ordinal, interval, or ratio-level variables and the dependent binary variable (Erguzel et al., 2019). Many researchers used RF, ANN, and LR to forecast groundwater pollution worldwide. ANN was used in China to predict geogenic groundwater F^- contamination across the country (Cao et al., 2022). It was also used to predict high NO_3^- in the groundwater of Harran Plain, Turkey (Yesilnacar et al., 2008). In the Yinchuan Region of central China, RF was employed to forecast NO_3^- pollution in groundwater (He et al., 2022). The RF method is used in Southern Spain to forecast NO_3^- in the groundwater and factors related to intrinsic and specific susceptibility (Rodriguez-Galiano et al., 2014). In Nigeria, research has demonstrated the appropriateness of utilizing ANN models for monitoring and evaluating water quality (Egbueri, 2021). ANN and multiple linear regression (MLR) exhibited strong reliability in monitoring groundwater resources. Both models demonstrated excellent performance, with MLR (ranging from 95 to 100%) outperforming ANN (ranging from 85 to 99%) in modeling the majority of potentially toxic elements (PTEs) and water quality indices (Agbasi & Egbueri, 2023). In the southern region of Nigeria, both MLR and multilayer perceptron neural networks (MLP-NN) methodologies were used to estimate and predict water quality indices, as well as the index of pollution (OIP) and water quality index (WQI). Remarkably minimal modeling errors were observed for both approaches, signifying the models' robust and concurrent predictive capabilities (Egbueri & Agbasi,

2022a). In the context of water quality analysis in Nigeria, the recent investigation synergistically integrated various soft computing algorithms. The outcomes validate that employing a combination of multiple models typically results in more robust and improved assessments compared to relying solely on an individual model (Egbueri & Agbasi, 2022b). However, these algorithms have been used independently to predict groundwater pollution, leaving gaps in determining the best ML model to predict groundwater contamination. The current study compares three machine learnings, RF, ANN, and LR, to predict the As in groundwater using binary classification analysis.

Pakistan is contending with a significant challenge of groundwater pollution caused by As, leading to adverse effects on groundwater quality across multiple regions, notably Punjab Province. The Pakistan Council of Research in Water Resources (PCRWR) identified elevated levels of As in Punjab's groundwater, surpassing the permissible drinking water limit set by WHO 2022 (2022). Moreover, in Sindh Province, the consumption of arsenic-contaminated drinking water has impacted 36% of the local population. Tragically, heavy metal contamination in the drinking water was responsible for the loss of 40 lives in the Hyderabad district in 2004 (Ullah et al., 2021). Groundwater pollution has increased due to the rapid population growth in the Indus plains of Punjab Province. Numerous studies have investigated groundwater As contamination across diverse settings, encompassing rural and urban areas and peri-urban zones of Pakistan (Fatima et al., 2018; Shahid et al., 2018a). As contamination in Vehari district's groundwater underscores its significance. Studies by (Shah et al., 2020) highlight elevated As levels, mainly attributed to geological factors and agricultural practices. These findings echo the concerns (Jat Baloch et al., 2022b) raised, highlighting the urgency of assessing health risks and implementing effective mitigation strategies.

A comprehensive analysis of potential drinking water contaminants remains imperative to safeguard the local population's well-being. Remarkably, minimal attention has been directed toward investigating As contamination and prediction within the drinking water sources of the Vehari district. Thus, this study assumes significance in pioneering: (i) an in-depth exploration of groundwater hydrogeochemistry,

emphasizing the spatial distribution of arsenic contamination; (ii) an innovative approach utilizing ANN, RF, and LR classifiers to unravel determinants influencing groundwater As; (iii) an assessment of the suitability of groundwater for human consumption through the Water Quality Index (WQI). The research innovation is particularly highlighted by its pioneering use of machine learning models, a previously unexplored approach in the study area. This utilization significantly improves the accuracy of arsenic prediction, leading to a substantial enhancement in our comprehension of local water safety.

Materials and methods

Study area

Vehari holds significance as a prominent district within the Punjab Province of Pakistan. Burewala,

Mailsi, and Vehari emerge within this district as key sub-districts (refer to Fig. 1). Geographically, the area is bounded by the Sutlej and Ravi rivers, positioned between coordinates 30°04'19" N and 72°35'28" E (Fig. 2). With a population of approximately three million, the Vehari district witnesses a climate characterized by scorching summers, temperatures peaking at 50 °C, and chilly winters, where temperatures can drop to about 5 °C. The summer also brings frequent dust storms, while the annual precipitation hovers around 125 mm. Groundwater is critical for agricultural and domestic needs in Vehari, Pakistan. Its accessibility promotes irrigation, which is essential to the local economy and ensures crop growth and food security. Groundwater also serves as a reliable buffer during droughts, protecting against erratic surface water availability. Significant human activities influencing groundwater quality and As levels in Vehari, Pakistan, include intensive agricultural practices involving fertilizer and pesticide use, industrial

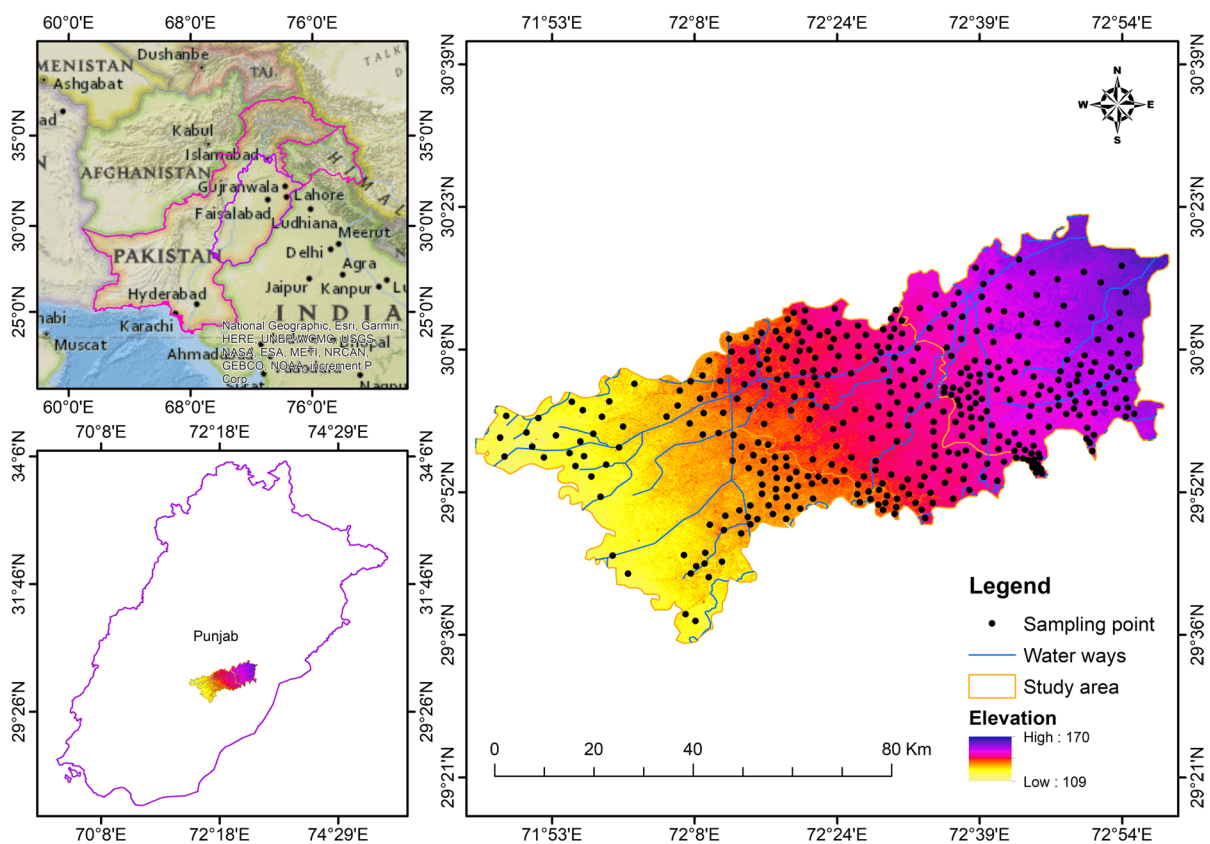
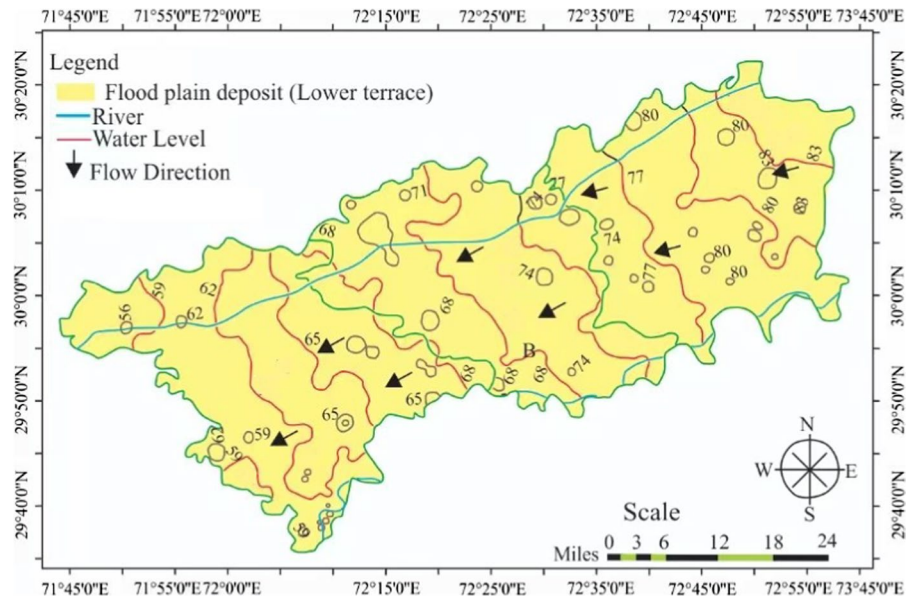


Fig. 1 Study area map showing the sampling location

Fig. 2 Hydrogeology map of the study area



operations with potential chemical releases, insufficient waste disposal practices, and possible contamination from unregulated domestic and municipal wastewater (Jat Baloch et al., 2022b).

Geology and hydrogeology

The geology and hydrogeology of the Vehari district reveal a strong relationship between borehole depths and fundamental aquifer properties, shedding light on subsurface characteristics and groundwater dynamics. The region is dominated by alluvial deposits, with borehole depths ranging from shallow to deep, revealing a stratigraphic succession of sediments such as silts, sands, and gravels. The Satluj and Ravi rivers run through the study area and serve as groundwater recharge sources (Khalid et al., 2020). The South Indus River forms the alluvial plain deposit, and its five major tributaries transport Pleistocene and Holocene sediments carried by the Ravi and Sutlej rivers (Fig. 2). The aquifer is made up of loose alluvial deposits that contain varying amounts of sand, a high percentage of fine sand and silt, and very little organic matter. Since the late Tertiary period, the Indus Rivers and streams have deposited these materials in the vast alluvial plain stretching from the Himalayan foothills to the Arabian Sea. The mineralogical evaluation identified aragonite, anhydrite, calcite, dolomite, gypsum, goethite, hematite, and halite as minerals

(Ahmad et al., 2002). During the Pleistocene epoch, the Indus River sediment deposits formed a substantial 400 m thick layer. The groundwater in the Punjab region is a mix of alluvial sand and alternating silt layers. The two main aquifer systems in the hydrogeological structure are the upper unconfined aquifer and the lower confined aquifer. Borehole data provides critical insights into aquifer depths, revealing that the upper aquifer is generally at shallower depths than the confined aquifer. The properties of aquifers are crucial to understanding the region’s hydrogeology. Aquifer yield is a measure of water provisioning capability, whereas storativity is a measure of water storage capacity. The rate of groundwater movement is significantly influenced by transmissivity, which measures the aquifer’s ability to transmit water (Shahid et al., 2018b). Borehole logs also reveal various lithological structures, such as fault zones and permeable layers, significantly impacting groundwater movement and distribution. The interaction of borehole depths, aquifer properties, and lithological structures shapes the groundwater flow regime (Ali et al., 2023).

Sampling and analysis

Four hundred groundwater samples were collected from the Vehari district. These groundwater samples were obtained explicitly from drinking wells at

diverse depths ranging from 50 to 400 feet (Table 2). All wells were flushed for at least 5 min to obtain fresh water before collecting groundwater samples. Groundwater samples (1000 mL each) were taken in duplicate in two separate plastic bottles having airtight caps. The samples were filtered through a 0.45 μm filter for further analysis. One water sample was acidified on-site by adding 2–3 drops of concentrated nitric acid (HNO_3) to stabilize As and metal ions and reduce precipitation (Shah et al., 2020). The acidified water samples were used to analyze total As contents and other elements. The second water sample was kept non-acidified to analyze various cations and anions. The American Public Health Association's recommended procedures were implemented (Jat Baloch et al., 2022b). Using a multi-parameter analyzer, the pH, electrical conductivity (EC), total dissolved solids (TDS), total hardness (TH), and turbidity of the study area were measured in situ (Hanna HI9829). The groundwater samples were then tested in the water quality laboratory of the Pakistan Council for Research in Water Resources (PCRWR) for further analysis. The samples were examined for significant anions such as NO_3^- , SO_4^{2-} , and PO_4^{3-} using a UV–VIS spectrophotometer. The concentration of F^- was determined using "Mohr's method and Fluoride Analyzer" ISE (ion-selective electrode) (Rashid et al., 2018a). The titration method was used to assess bicarbonate (HCO_3^-) and chloride (Cl^-). Volumetric titration with ethylene diamine tetra acetic acid was utilized to determine calcium (Ca^{2+}) and magnesium (Mg^{2+}) concentrations. The sodium (Na^+) and potassium (K^+) concentrations were measured using a flame photometer (Zhou et al., 2021). As levels in the samples were measured using an atomic absorption spectrophotometer (AAS Vario 6, Analytik Jena, Jena, Germany (Baloch et al., 2022)). The charge balance error (CBE) for each sample was calculated (ionic concentrations are measured in meq/L) to ensure the accuracy of the results. Groundwater samples containing $\pm 5\%$ CBE were chosen for further examination (Jat Baloch et al., 2022a).

Statistical and hydrochemical analysis

Statistical software XL STAT 2021 was employed to compute the mean values, including minimum, maximum, average, and standard deviation, for each parameter. Piper diagram was utilized to determine

the hydrogeochemical type and concentration of major anions and cations in the water samples and identify geochemical processes that contribute to assessing groundwater quality (Ullah et al., 2022b). The Piper diagram was produced using Grapher, and the Gibbs diagram was used to determine groundwater evolution. Saturation indices were calculated using PHREEQC Interactive to measure water's mineral balance and dissolved mineral reactivity.

Preprocessing of data for machine learning model

The input parameters were EC, pH, TDS, Turbidity, Hardness, Cl^- , HCO_3^- , Ca^{2+} , Mg^{2+} , SO_4^{2+} , K^+ , Na^+ , Fe^{2+} , NO_3^- , F^- , and the dependent variable (As). All As concentrations less than 10 $\mu\text{g/L}$ were assigned a value of zero (0), and concentrations greater than 10 $\mu\text{g/L}$ were given one (1) value. To improve the model's speed and accuracy, the independent variables for the three algorithms were scaled between 0 and 1 (Nafouanti et al., 2021). Subsequently, the dataset was randomly partitioned into two segments: 80% designated for the training phase and 20% allocated for testing. The adjustment of actual groundwater variable concentrations, particularly for As concentrations, serves a scientific purpose in enhancing the modeling process. By categorizing As concentrations as below 10 $\mu\text{g/L}$ (assigned as 0) or above 10 $\mu\text{g/L}$ (assigned as 1), the study aims to create a binary classification framework that aligns with regulatory thresholds for safe drinking water. This approach offers several benefits: it simplifies the modeling task, focusing on classifying water as safe or contaminated and addressing potential noise and variability in the dataset. Moreover, it aligns with real-world decision-making scenarios where the primary concern is identifying water sources with elevated As levels that exceed permissible limits. This categorization facilitates efficient model training, convergence, and prediction accuracy, contributing to a more practical and actionable outcome for groundwater quality assessment and management strategies.

Choosing the appropriate input

Feature selection is crucial in classification because it enhances the classifier's performance while reducing computation complexity by eliminating duplicated data (Zebari et al., 2020). In this study, filter methods were

used to select the relevant inputs. These approaches are faster than wrapper methods because they do not require model training. They can also link the independent and dependent variables (Coulibaly et al., 2000). The chi-squared method can create independent comparison tests (Zebari et al., 2020). For feature selection, chi-squared analysis was used to compute the chi-squared score of each class, resulting in a ranking list of all features. The numeric attributes were discretized to use the chi-squared statistic to find inconsistencies in the data (Kim, 2017). The following equation was used to calculate a feature's chi-squared score.

$$X^2 = \sum_{i=1}^r \sum_{j=1}^{c'} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{1}$$

where *c* represents the total classes, and *r* denotes the discrete intervals for the specified feature. *n_{ij}* signifies the observed frequency of the groundwater samples in the *i*th interval and *j*th class.

If *n_i = c_j = 1*, the number of samples in *i*th interval for a feature is *n_{ij}*; otherwise, the number of samples in the *i*th interval for a feature is *n_j = r_i = 1*. The sample count for class *j* is *n_j*, the total sample count is *n*, and the expected frequency of *n_{ij}* is $\mu_{ij} = \frac{n_{i.}n_{.j}}{n}$.

When the observed number is close to the expected number, and the Chi-squared value is small, the variables are considered independent. Because of the higher Chi-Squared value, a variable is significant to the outcome and should be used to train models. Python's sklearn module and the "SelectkBest" function were used to select the variables, which kept the first *k* (no of the samples being summed) input variables (Table 1). The twelve (12) variables with high Chi-squared values were chosen as critical groundwater inputs. pH, TDS, SO₄²⁻, Na⁺, Fe²⁺, Cl⁻, HCO₃⁻, Ca²⁺, Mg²⁺, NO₃⁻, K⁺ and F⁻ for the As prediction.

Artificial neural networks (ANN)

Artificial neural networks has proven to be an effective categorization, clustering, pattern recognition, and prediction model. ANN is an ML model that

outperforms conventional regression and statistical models (Musa et al., 2019). ANN are multilayered biologically inspired computer models with input, hidden, and output layers. The primary processing unit of ANN is the neuron, which connects all layers (Afzaal et al., 2019). Multilayer perceptron (MLP) neural networks used in this study are among the most common types of ANN. MLP includes an input layer with source neurons, one or more hidden layers of neurons, and an output layer. The number of nodes in the input and output layers changed according to the number of input and output variables (Fig. 3). The generalization potential of the network is determined by the number of hidden layers and the number of nodes per hidden layer and it contains two layers. The relatively limited number of hidden layers and neurons may cause underperformance.

In contrast, too many hidden nodes may overfit training data and poorly generalize new input (Otchere et al., 2021). In this study, the "adam" optimizer was utilized to update the weight in the network. The permutation feature has been used to identify critical variables in the correlation of predictors and dependent variables. It describes the impact of variable elimination on network accuracy.

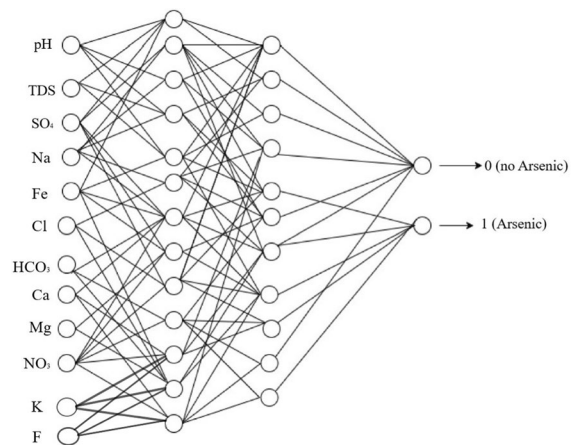


Fig. 3 Artificial Neural Network structure with the Inputs variables

Table 1 Selection of Relevant Inputs by using the Chi-Squared analysis

Variables	pH	TDS	SO ₄ ²⁻	Na ⁺	Fe ²⁺	Cl ⁻	HCO ₃ ⁻	Ca ²⁺	Mg ²⁺	NO ₃ ⁻	K ⁺	F ⁻
Score	92,626.6	9185.3	5882.9	4644.4	3035.6	2503.5	1195.9	958.3	651.1	318.6	181.6	8.3

Random forest modeling

Random forest avoids the limitations of overfitting and instability when only one decision tree is used. The primary goal of RF is to generate many decision trees from random subsets of the original training dataset. The average forecasts of these single trees are used to increase the model's generalization (Wu et al., 2020). RF classification was used in this research to predict As pollution in groundwater. To generate a training subset for every tree, a bootstrapping technique determines the training dataset into an "in-bag" subset for decision tree training and an "out-of-bag (oob)" subset that is not used in the training process. Internal validation is performed because each tree is partitioned. The mean of all oob forecasts would provide a metric for the accuracy rate of the RF model, and oob samples from each tree are used to assess its efficiency. A decision tree's in-bag and out-of-bag sample sizes are 66.67 and 33.333% (2:1) of the original training dataset. After the model has been formed and fitted with the training dataset, its performance is assessed using the test dataset. Consequently, upon both training and test sets, the model makes oob predictions. To evaluate model performance, metrics including mean absolute error, root mean square error, and the coefficient of determination (R^2) are employed to measure the disparities between observed and predicted response variables (Markwart et al., 2019). In addition, the trained and validated RF model evaluates predictor variable significance to determine how each predictor factor influences the response variable. The RF model in the current study was built using 100 trees. RF can identify and characterize the critical predictive variables that cause groundwater contamination. The permutation function was utilized to determine the significant factors in the association between predictors and dependent variables (Hussain et al., 2021). A considerable decline in impurity constitutes an essential split. The greater the significance of the variable, the more significant the mean impurity reduction.

Logistic regression

Logistic regression (LR) is a statistical model that uses a logistic function to illustrate a binary dependent variable. LR is a method for defining the requirements of a logistic model in regression analysis

(Wasserman & Pattison, 1996). This study uses LR to forecast the level of As in groundwater. The LR equation is as follows:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (2)$$

β_0 and β_1 are estimated parameters.

Machine learning model assessment criteria

Data analysis in a confusion matrix is a standard method for assessing predictive performance. The accuracy, specificity, sensitivity, and error have been computed to determine the model prediction. AUC (Area Under the Curve) is a metric commonly used in binary classification to assess the performance of machine learning models. It represents the area beneath the receiver operating characteristic (ROC) curve, reflecting the model's ability to distinguish between positive and negative classes. A higher AUC value (closer to 1) indicates better model discrimination and classification accuracy. LR was also evaluated using the ROC and AUC. A confusion matrix determines the ability to predict binary classification correctly and accurately. It demonstrates how the model distinguished predicted and actual values (Nafouanti et al., 2021). To analyze the classified data percentages, the prediction has been evaluated by comparing them to the identified concentration. The sensitivity is the percentage of As correctly classified, while the specificity is the percentage of non-arsenic correctly classified. The Python 3.9 programming language was used to create the three models.

The confusion matrix metrics equation is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

$$\text{Error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

where TP=True Positive, TN=True Negative. FP=False Positive, FN=False Negative.

Water quality index

The study employed the WQI to evaluate the suitability of groundwater for drinking purposes in the study area, as conducted by (Agbasi & Egbueri, 2023; Egbueri & Agbasi, 2022a; Omeka & Egbueri, 2023; Onyemesili et al., 2022). The WQI values were calculated based on the World Health Organization (WHO, 2022) drinking water standards for nine parameters, TDS, pH, Turbidity, Ca²⁺, Mg²⁺, Na⁺, K⁺, Cl⁻, SO₄²⁻, HCO₃⁻, NO₃⁻, F⁻, and Fe²⁺. To calculate the WQI, three computing steps were undertaken. First, weights (*w_i*) were assigned to each parameter based on their significance in determining groundwater quality, with Table 2 providing the weight and relative weight of all hydrochemical parameters.

The relative weight (*W_i*) for each parameter was computed using Eq. (7), where *W_i* denotes the relative weight, *w_i* represents the weight of the specific parameter, and *n* indicates the total number of parameters. This step aimed to weigh the importance of each parameter proportionally.

$$W_i = \frac{W_i}{\sum_{i=1}^n W_i} \tag{7}$$

The second step involved determining each parameter’s quality rating scale (*q_i*) using Eq. (8). In this equation, *q_i* represents the quality ranking, *C_i* signifies the parameter’s quality in milligrams per liter (mg/L), and *S_i* represents the WHO (2022) standard for that parameter. This calculation allowed for the evaluation of the quality of each parameter with the established standards.

$$q_i = \frac{C_i}{S_i} \tag{8}$$

The sub-index (*SI_i*) for each parameter was computed using Eq. (9) to consolidate the various parameter sub-indices into a single representative value. In this equation, *SI_i* represents the sub-index of the *i*th parameter, *W_i* signifies the relative weight of that parameter, and *q_i* corresponds to the rating associated with the concentration of the specific parameter. This step aimed to reflect the significance of each parameter in contributing to the overall assessment.

$$SI_i = W_i \times q_i \tag{9}$$

Finally, the comprehensive WQI was determined by summing up all the individual sub-indices using Eq. (10). This final index provided a holistic understanding of the drinking water quality in the research area, aligning with the WHO 2022 drinking water quality standards for the specified hydrochemical parameters.

Table 2 The weight and relative weight of hydrochemical parameters

Chemical parameters	WHO 2022 standards (mg/L) (<i>S_i</i>)	Weight (<i>w_i</i>)	Relative weight (<i>W_i</i>)
TDS	1000	1	0.038461538
pH	6.5–8.5	1	0.038461538
Turbidity	5	2	0.076923077
Ca ²⁺	200	1	0.038461538
Mg ²⁺	150	1	0.038461538
Na ⁺	200	2	0.076923077
K ⁺	12	3	0.115384615
Cl ⁻	250	2	0.076923077
SO ₄ ²⁻	250	2	0.076923077
HCO ₃ ⁻	250	4	0.153846154
NO ₃ ⁻	10	1	0.038461538
F ⁻	1.5	5	0.192307692
Fe ²⁺	0.3	5	0.192307692
		Sum of <i>w_i</i> = 30	Sum of <i>W_i</i> = 1

$$WQI = \sum_{i=1}^n SI_i \quad (10)$$

Results and discussion

Hydrogeochemical analysis of groundwater

The hydrogeochemical characteristics of the groundwater samples are displayed in Table 3 and compared with the WHO 2022 standards for drinking water quality (Organization, 2022). The EC shows the ability of water to transmit an electric current between dissolved salts. However, EC ranges from 85 to 4550 $\mu\text{S}/\text{cm}$ with a mean value of 1363.10, showing that groundwater mineralization is responsible for elevated EC saturating salinity in the groundwater system. The findings suggest that the groundwater chemistry in the study area is impacted by geochemical processes, rock-water interactions, and human activities (Adimalla et al., 2021). Total dissolved solids (TDS) measurements are essential for reporting dissolved chemical concentrations. TDS concentrations varied from 234 to 3173 mg/L, with a mean of 968.20 mg/L. The elevated TDS is due to salt leaching and sewage infiltration (Khan et al.,

2018). Higher salinity in groundwater cause high EC and TDS levels, typically related to semi-arid and arid climatic conditions (Herczeg et al., 2001). The groundwater pH varies from 6.78 to 8.18, with a mean value of 7.16, indicating slightly alkaline. Because of pH variations, the chemical composition of groundwater changes, and this variation is primarily determined by lithology. Weathering and chemical reactions of plagioclase feldspar in sedimentary rocks (Ali et al., 2023). The groundwater Turbidity levels ranged from 0.3 to 188.0 NTUs (Nephelometric Turbidity Units), with a mean of 6.34 NTU. Poorly constructed and too-shallow wells can cause high turbidity (Azis, 2015). Furthermore, the alkaline condition increases conductivity over time due to the dissolution process. The total Hardness ranged between 100 and 820 mg/L, with an average of 361.39 mg/L. Water with a hardness of > 500 mg/L is unsafe for human consumption (WHO, 2022). The amount of CO_2 in the soil increases due to humus decomposition and respiration in the topsoil. The breakdown of feldspar and carbonate minerals is accelerated by high soil CO_2 , resulting in high groundwater alkalinity (Roy et al., 2018).

The groundwater is dominated by cations in the following order: $\text{Na}^+ > \text{Ca}^{2+} > \text{Mg}^{2+} > \text{K}^+ > \text{Fe}^{2+}$. The concentration of Na^+ ranged from 13 to 662 mg/L, with an average of 148.73 mg/L. The

Table 3 Statistical analysis of physicochemical parameters for groundwater samples for the study area

Parameters	Minimum	Maximum	Mean	SD	WHO (2022) Standard
EC($\mu\text{S}/\text{cm}$)	85	4550	1363.10	748.009	1000
TDS (mg/L)	234	3173	968.20	469.715	1000
pH	6.780	8.180	7.16765	0.277313	6.5–8.5
Turbidity	0.30	188.00	6.3417	22.93979	5
TH	100	820	361.39	126.042	300
HCO_3^- (mg/L)	80	900	319.90	104.604	250
Cl^- (mg/L)	10	518	92.48	74.187	250
SO_4^{2-} (mg/L)	18	1432	255.02	178.978	250
Ca^{2+} (mg/L)	8	208	86.20	34.592	200
Mg^{2+} (mg/L)	2	104	35.25	16.511	150
Na^+ (mg/L)	13	662	148.73	118.331	200
K^+ (mg/L)	2.6	74.0	9.072	7.2372	12
NO_3^- (mg/L)	0.01	17.66	1.7610	3.14160	50
Fe^{2+} (mg/L)	0.01	3.92	1.1455	19.59425	0.3
F^- (mg/L)	0.00	3.15	0.5176	0.29004	1.5
As (ug/L)	0.1	52	11.64	11.393	10

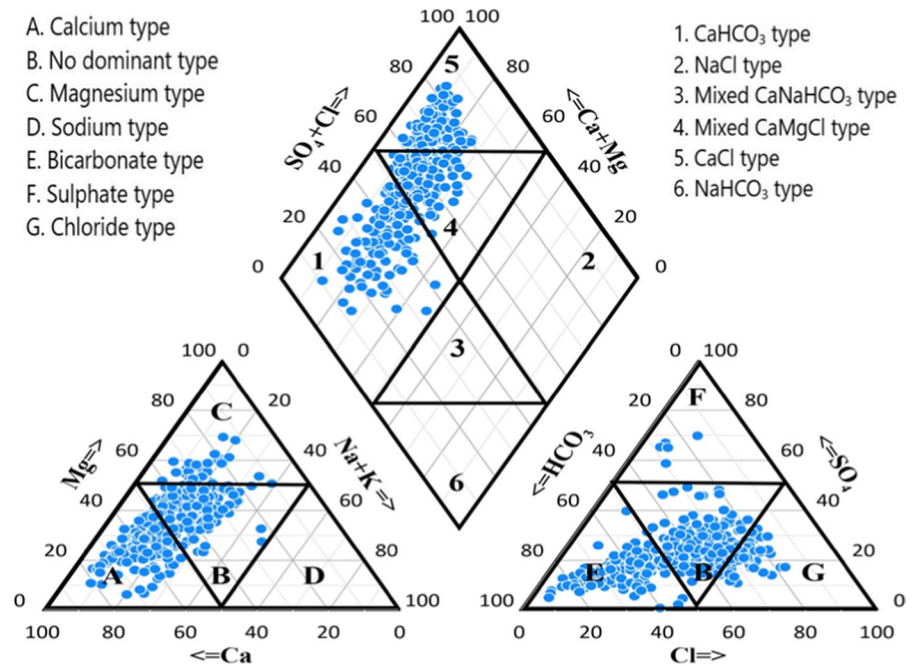
high levels of Na^+ in the groundwater are attributed to ion exchange caused by silicate weathering, saline water infiltration, or clay minerals (Mitchell et al., 2018). Furthermore, agricultural activities in the research region may impact the prevalence of Na^+ in groundwater. The average concentration of Ca^{2+} was 86.20 mg/L, with a range of 8–208 mg/L. Higher Ca^{2+} content is from geological sources, such as the dissolution of carbonate and evaporite minerals or carbonate minerals within rock formations (Chidambaram et al., 2018). Mg^{2+} levels varied between 2 and 104 mg/L, with a mean value of 35.25 mg/L. The elevated Mg^{2+} originating from minerals like mica, gypsum, and dolomite could also arise through ion exchange (Chidambaram et al., 2018). Groundwater with higher Ca^{2+} and Mg^{2+} concentrations is classified as hard water. The K^+ concentrations in the study region ranged from 2.6 to 74.0 mg/L, with an average of 9.072 mg/L, and were influenced by agricultural activities and water seepages from agrarian lands. Natural sources of K^+ ions, such as silicate minerals, also contribute to K^+ ions in groundwater. The maximum permissible Fe^{2+} concentration in groundwater is 0.3 mg/L, according to the (WHO, 2022). The Fe^{2+} levels in the study region ranged from 0.01 to 3.92 mg/L, with an average of 1.1455 mg/L, and were primarily sourced from ferruginous minerals on the Earth's surface (Raju, 2006).

The dominant anions in the groundwater are $\text{HCO}_3^- > \text{SO}_4^{2-} > \text{Cl}^- > \text{NO}_3^- > \text{F}^-$. The concentrations of HCO_3^- ranged from 80 to 900 mg/L, with an average of 319.90 mg/L, making it the most prominent anion. The presence of HCO_3^- in groundwater can be attributed to the breakdown of carbonate minerals and the interaction of atmospheric CO_2 with silicate minerals (Fornes et al., 2020). The SO_4^{2-} concentrations varied from 18 to 1432 mg/L with a mean value of 255.02 mg/L. The higher SO_4^{2-} levels in groundwater resulted from agricultural activities (Manjusree et al., 2009). Groundwater Cl^- concentrations ranged from 10 to 518 mg/L, with a mean value of 92.48 mg/L. Higher Cl^- content in the aquifers is caused by saline water infiltration and evaporite dissolution (Gopinath et al., 2018). The NO_3^- varied from 0.01 to 17.66 mg/L with a mean value of 1.76 mg/L. Fertilizer runoff, septic systems, and improperly treated wastewater are the anthropogenic sources of NO_3^- (Selvakumar et al., 2017). The As concentrations in the study

area ranged from 0.1 to 52 $\mu\text{g/L}$, with an average of 11.64 $\mu\text{g/L}$. Elevated levels of As in groundwater is due to natural and anthropogenic sources (Adimalla et al., 2018). The higher levels of HCO_3^- in the groundwater of Vehari district result in increased As concentration showing an oxidative condition in the aquifers. Most regions in southern Punjab contain high levels As in groundwater due to arsenic minerals, making most of the water resources unsuitable for drinking. Vehari district faces a critical challenge due to the widespread As contamination (Ullah et al., 2021).

Hydrogeochemical evolutionary processes

The hydrochemical facies diagram depicts the groundwater interactions in a lithological formation (Boateng et al., 2016). The chemical differences among the groundwater samples are shown in the Piper diagram (Fig. 4). Most of the samples fall into Zones 1 (CaHCO_3 type), 4 (mixed CaMgCl type), and 5 (CaCl) and no dominance type, indicating that the rock-water interaction plays a significant role in determining the hydrochemical composition. Zone 1 (CaHCO_3 type) represents fresh recharge water samples. Regarding cations, the groundwater samples can be classified into Zone B (mixed type) or Zone D ($\text{Na} + \text{K}$ type), highlighting the significance of silicate weathering and ion exchange. The majority of the groundwater samples are classified into Zone B (mixed type) and E (HCO_3^- type) from the anions' perspective, with a few samples falling into Zone G (Cl type). This implies that carbonate weathering and evaporite dissolution are the dominant processes, whereas gypsum dissolution is negligible in the study area. For the cations, the majority of the samples fall into Zone B (No dominance), Zone A (Calcium type), and Zone C (Magnesium type). For Ca^{2+} and Mg^{2+} components in water samples, limestone and sandstone weathering significantly influences the groundwater system (Mallick et al., 2021). In the Piper plot, the types of waters Na^+ , SO_4^{2-} , Ca^{2+} , and Mg^{2+} were demonstrated As released by sedimentary rocks into groundwater. As mobilization in groundwater is caused by several vital mechanisms, including calcium dissolution, salt mineral dissolution, and desorption (Jat Baloch et al., 2020).

Fig. 4 Geochemical evolution of groundwater types

Gibbs diagram

The Gibbs diagram portrays the groundwater chemistry-influencing variables: evaporation dominance, precipitation dominance, and weathering dominance (Jat Baloch et al., 2021; Rashid et al., 2018b; Salem et al., 2015). Most of the samples are plotted in the rock dominance region in Fig. 5, signifying that rock dominance influences the majority of groundwater, while a few are also in the evaporation dominance. Rock weathering is the foremost driving force behind the heightened presence of minerals within the groundwater system. This enrichment is facilitated by intermingling soluble salts and minerals within the groundwater. Furthermore, the extended period of water–rock interaction, resulting from the prolonged residence time, allows the potential dissolution of minerals. This phenomenon underscores the complex interplay between geological processes and groundwater composition, a pivotal focus in scientific research (Tariq et al., 2022).

Pearson correlation

The findings of Pearson's correlation analysis are displayed in Table 4. In the conventional interpretation, quality parameters exhibiting correlation

coefficients (r) of <0.5 , between 0.7 and 0.5, and >0.7 signify weak, moderate, and strong relationships, respectively (Onyemesili et al., 2022). From the correlation matrices for Vehari, we were able to understand the geochemical process in the study area. The strong correlations between EC and TDS, HCO₃⁻, Cl⁻, SO₄²⁺, Mg²⁺, Na⁺, TH, and F⁻ indicate higher ion exchange possibilities in the aquifers. The significant correlation between TDS and HCO₃⁻, Cl⁻, SO₄²⁺, Ca²⁺, Mg²⁺, Na⁺, and TH. As the TDS value increase, all ionic concentrations also increase, primarily due to weathering of sedimentary rocks. TH positively correlated with EC, TDS, HCO₃⁻, Cl⁻, SO₄²⁺, and Ca²⁺, illustrating that groundwater has elevated hardness due to Ca²⁺ and Mg²⁺, and other ions in the study area (Xue-Jie et al., 2013). The significant correlation between HCO₃⁻ with Cl⁻, SO₄²⁺, Mg²⁺, and Na⁺ suggests a significant contribution from multiple anthropogenic sources like improper disposal of wastes, agricultural activity, sanitation, discharge of industrial effluents, and organic decomposition in the study area. In Vehari, As exhibited a negative correlation with EC, TDS, pH, Turbidity, HCO₃⁻, Cl⁻, Mg²⁺, Na⁺, K⁺, Hardness, and NO₃⁻, Fe²⁺ and F⁻. Such correlations highlighted the influence of pH on As concentration in groundwater (Jia et al., 2023).

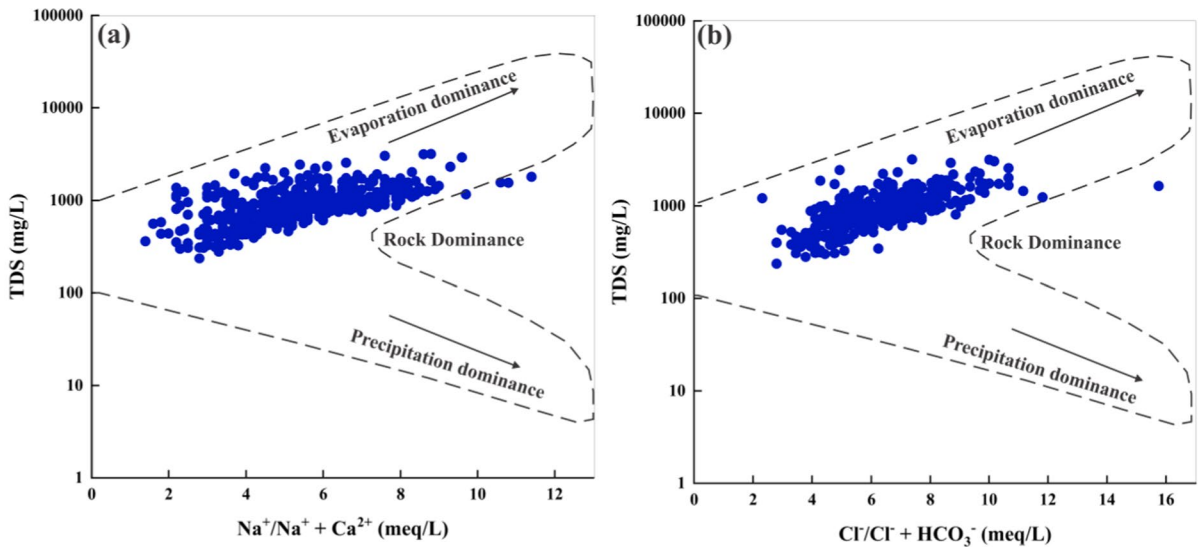


Fig. 5 The Gibbs diagram demonstrates the ionic composition of the samples of groundwater **a** Na/Na+Ca mg/L versus Log TDS, **b** Cl/Cl+HCO₃ mg/L versus Log TDS

Principal component analysis (PCA)

Principal component analysis was implied to find and classify the sources that influenced the groundwater variables. The factors related to groundwater were subjected to a PCA, as shown in Table 5. The varimax rotation was applied to the PCA results to understand better the factors that impact groundwater (Rashid et al., 2020; Zhang et al., 2020). Four components were obtained, with eigenvalues of 6.476, 1.874, 1.359, and 1.016, accounting for 38.096, 11.026, 7.993, and 5.974% of the total variability, respectively (Fig. 6). PC1 had 38.89% of variability with an eigenvalue of 6.476. The significant loadings factors of EC, TDS, HCO³⁻, Cl⁻, SO₄⁻², Mg²⁺, Na⁺, and TH were calculated to be 0.959, 0.980, 0.763, 0.865, 0.911, 0.718, 0.868, and 0.723. Thus, PC1 showed the highest contribution of strong loading factors in PCA results, demonstrating the geogenic and anthropogenic sources in the study area. The PC1 results indicate the ionic formation in groundwater, resulting from the ion exchange process, dissolution of minerals, and weathering of rocks. PC2 exhibits 11.026% variability with an eigenvalue of 1.874. The moderate loadings factors of groundwater variables were Ca²⁺ (0.731) and TH (0.645) in the study area. The levels of Ca²⁺ and Hardness are likely to be influenced by anthropogenic and weathering activities (Li

et al., 2020a). The PC3 and PC4 showed the lowest contribution in PCA results with 7.993 and 5.974% variability and eigenvalues of 1.359 and 1.016. The moderate factors of variables in PC3 and P4 could be associated with anthropogenic activities leading to influence the hydrochemical characterization of groundwater aquifers. These results highlight the contribution of anthropogenic and natural factors to groundwater contamination in the area under study region.

Machine learning model evaluation and comparison

The test predictor data were used to evaluate the models' precision in predicting the presence of As in groundwater after model development and training. The ANN, RF, and LR evaluation metrics were obtained from their confusion matrix. Tables S1, S2, and S3 provide more information. Based on the assessment criteria applied to the three models, the RF model demonstrated accuracy, error rate, specificity, and sensitivity values of 0.85, 0.10, 0.79, and 0.95, respectively (Table 6). High sensitivity over specificity means fewer false negatives in binary classification, indicating a good prediction model. RF's capability to forecast groundwater pollution for F⁻ has previously been investigated, which supports our study (Nafouanti et al., 2021). The accuracy of

Table 4 Pearson correlation matrix of the study area's physiochemical parameters for groundwater

	EC	TDS	pH	Turbidity	HCO ₃	Cl	SO ₄	Ca	Mg	Na	K	TH	NO ₃	Fe	F	As
EC	1															
TDS	0.942**	1														
pH	-0.006	-0.018	1													
Turbidity	-0.038	-0.027	-0.07	1												
HCO ₃	0.715**	0.673**	-0.01	-0.039	1											
Cl	0.877**	0.838**	0.118*	-0.023	0.556**	1										
SO ₄	0.869**	0.896**	-0.07	-0.019	0.551**	0.731**	1									
Ca	0.348**	0.560**	-0.148**	0.024	0.313**	0.306**	0.406**	1								
Mg	0.599**	0.674**	0.071	-0.021	0.525**	0.536**	0.327**	0.086	0.390**	1						
Na	0.872**	0.764**	0.031	-0.052	0.685**	0.782**	0.738**	0.086	0.264**	0.186**	1					
K	0.273**	0.268**	0.06	0.076	0.288**	0.259**	0.205**	0.154**	0.264**	0.186**	0.245**	1				
TH	0.555**	0.740**	-0.06	0.003	0.495**	0.507**	0.561**	0.854**	0.758**	0.265**	0.245**	0.203**	1			
NO ₃	0.137**	0.142**	-0	-0.085	0.172**	0.155**	0.061	0.132**	0.212**	0.092	0.170**	0.203**	0.203**	1		
Fe	0.112*	0.111*	-0.02	-0.012	0.130**	0.058	0.148**	-0.01	0.052	-0.05	-0.01	0.02	-0.03	0.08	1	
F	0.503**	0.380**	0.041	-0.046	0.353**	0.398**	0.434**	-0.188**	0.186**	0.574**	0.116**	-0.03	0.07	0.08	0.08	1
As	-0.029	-0.013	-0.03	-0.067	-0.102*	-0.049	0.007	0.065	-0.170**	-0.063	-0.115*	-0.04	-0.195**	-0.05	-0.05	1

Correlation is significant at the 0.01 level (2-tailed). Bold = strong correlation (> 0.90)

Table 5 Principal component analysis of selected groundwater parameters for Vehari district

Component	F1	F2	F3	F4
EC	0.959	- 0.180	0.071	- 0.048
TDS	0.980	0.054	0.122	- 0.019
pH	- 0.087	- 0.073	0.047	- 0.402
Turbidity	0.022	0.240	0.382	0.299
Alkalinity	0.098	- 0.205	0.018	0.467
HCO ₃	0.763	- 0.075	- 0.115	0.123
Cl	0.865	- 0.138	0.016	- 0.111
SO ₄	0.911	- 0.128	0.206	- 0.027
Ca	0.497	0.731	0.187	0.034
Mg	0.718	0.266	- 0.215	- 0.005
Na	0.868	- 0.413	0.109	- 0.014
K	0.339	0.115	- 0.396	- 0.121
TH	0.723	0.645	0.013	0.020
NO ₃	0.206	0.174	- 0.636	- 0.108
Fe	0.128	- 0.168	0.031	0.606
F	0.450	- 0.655	- 0.091	- 0.107
As	- 0.069	0.001	0.686	- 0.331
Eigenvalue	6.476	1.874	1.359	1.016
Variability (%)	38.096	11.026	7.993	5.974
Cumulative %	38.096	49.122	57.115	63.088

Extraction method: principal component analysis. Rotation method: Kaiser normalization of Varimax. Bold values show higher loading value

RF efficiency is improved in this work by identifying appropriate inputs and employing many trees, resulting in a performance boost for the model. The ANN’s accuracy, error rate, specificity, and sensitivity were 0.80, 0.20, 0.73, and 0.88, respectively. The current finding supports the previous research, Water quality indicator forecasting for irrigation applications using ANN (Abrahart et al., 2005; El Bilali et al., 2021) and (Awu et al., 2015). In this study, increasing the number of hidden layers in the network training improved ANN performance. By increasing the number of hidden layers, accuracy can be significantly enhanced (Karsoliya, 2012). In ANN, a suitable number for network training with two hidden layers can be obtained.

The LR’s accuracy, error rate, specificity, and sensitivity were 0.59, 0.41, 0.52, and 0.63, respectively; the model’s capabilities were assessed using the ROC curve (AUC) (Fig. 7). LR’s AUC was 0.73; the current finding supports the findings of previous research, Groundwater NO₃ pollution in a semi-arid environment utilizing integrated parametric IPNOA and data-driven logistic regression (Rizeei et al., 2018). The diminished effectiveness of the ANN model, in contrast to the RF model, arises from the ANN model’s limitation in making predictions outside the range of its training data. Consequently, the intricate challenge of overfitting becomes pronounced within the ANN training data (Al-Mukhtar, 2019). Because the RF model avoids overfitting and combines many trees to generate a prediction. Regarding accuracy, specificity, and sensitivity,

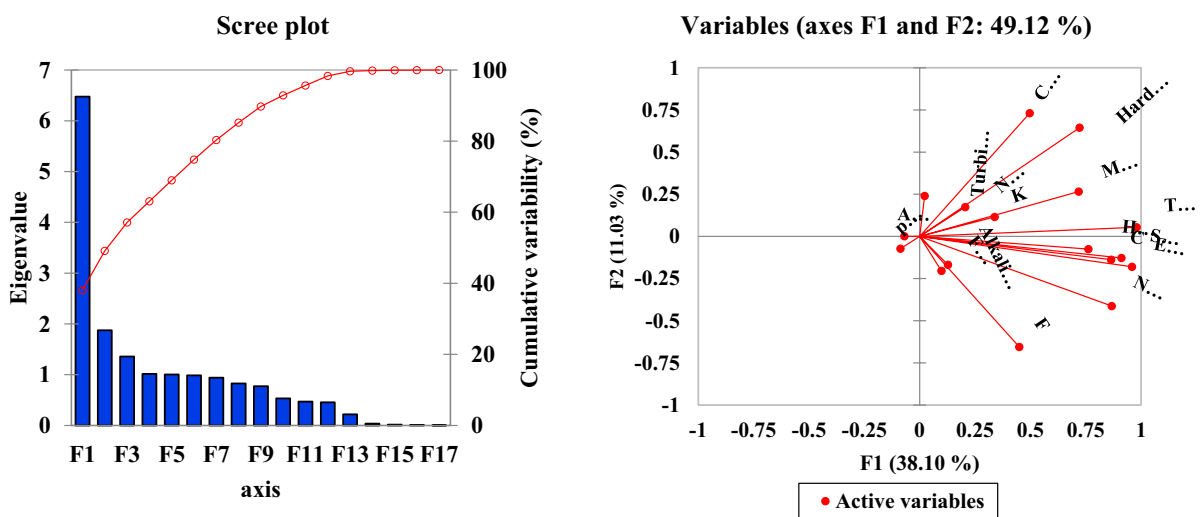
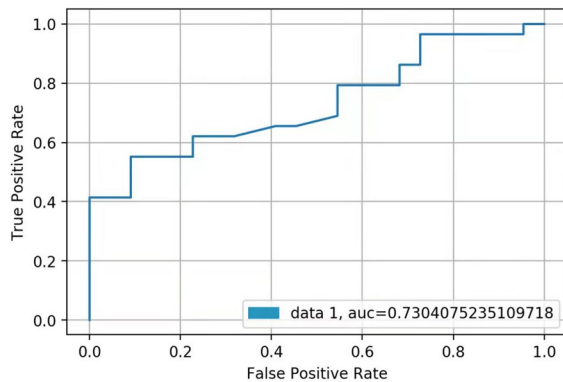


Fig. 6 Sum of all the calculated factors, b Contribution of the four loading factors F1, F2, F3 and F4 after varimax rotation

Table 6 Performance for Logistic Regression using ROC (AUC) curve

Metrics	RF	ANN	LR
Accuracy	0.85	0.80	0.59
Error rate	0.10	0.20	0.41
Specificity	0.79	0.73	0.52
Sensitivity	0.95	0.88	0.63

**Fig. 7** Performance for Logistic Regression using ROC (AUC) curve

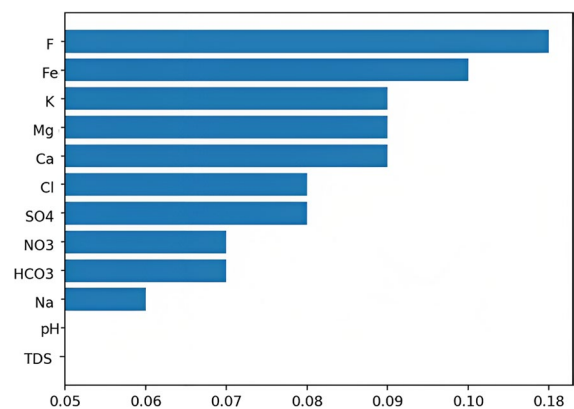
the LR model performed the worst of the three models (Table 6). Low-dimension data in the training data set can reduce LR performance. The model on the test data set may be overfitting and incorrect. Despite their poor performance in the current study, ANN and LR have advantages when forecasting groundwater contamination in previous studies. Because of the presence of numerous variables, the process of groundwater pollution is difficult to comprehend. Hence, the model's precision and dependability increase proportionally with the algorithm's enhanced adaptability (Tsoar et al., 2007). An algorithm's structure, the data type, and the parameter selections influence its performance (Üstün et al., 2005). In this classification task, feature selection should be considered for statistical analysis to produce an excellent predictive model.

Identifying the variables that impact arsenic mobilization

The mean decrease in impurity (MDI), a factor significance metric via RF, was used to determine the relationship between predictors and As (He et al.,

2022). It's a tree-specific variable importance metric calculated with Python's RF "skirt-learn" module's feature importance implementation. Each time a variable is chosen to split a node, the cumulative MDI per feature across all forest trees is calculated. Factors dividing nodes closer to the tree root have a higher significance value (Fig. 8). The F^- , Fe^{2+} , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , SO_4^{2-} , NO_3^- , HCO_3^- , and Na^+ variables are at the top of the plot with the highest MDI values. The plot shows that pH and TDS have a low MDI in the research area.

The MDI was previously used to find the essential components in data that influence the dependent variable (Bylander, 2002). Furthermore, the MDI identified significant variables related to the dependent variable in microarray and facies estimation. It was employed to find essential predictors of the dependent variables (Bhattacharya & Mishra, 2018). According to the MDI results, the variables influencing As in the study region are F^- , Fe^{2+} , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , SO_4^{2-} , NO_3^- , HCO_3^- , and Na^+ , which is consistent with previous research findings (Jat Baloch et al., 2022b; Rashid et al., 2018a; Tahir & Rasheed, 2013). By evaluating the variable importance of ANN, the permutation feature was utilized to find the utmost influential aspects of the output. When a single variable is removed, the permutation lowers the final model score (Chae et al., 2016). Twelve (12) networks were tested to discover the most significant factors in the outcome. After removing a variable, each showed a change in network accuracy variance (Table 7).

**Fig. 8** Important Features to the Arsenic using Mean Decrease in Impurity in Random Forest

The accuracy is 0.80 after omitting the pH and TDS, the same as the original model accuracy. Consequently, the potential exclusion of pH and TDS from the model arises, given their limited impact on network accuracy. This observation underscores that pH and TDS insignificantly influence the concentration of As within the study region. In contrast, when additional variables such as F^- , Fe^{2+} , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , SO_4^{2-} , NO_3^- , HCO_3^- , and Na^+ are removed from the model, the model's accuracy decreases, showing their significance to the As model. Permutation was previously used in research to identify critical components in dissolved oxygen (DO) (Matayoshi et al., 2019). In the current research, the permutation feature and the MDI give similar outcomes to the variables affecting As in the study region, F^- , Fe^{2+} , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , SO_4^{2-} , NO_3^- , HCO_3^- , and Na^+ . When analyzing the correlation between the input and output variables, the permutation feature outperforms the MDI feature. The permutation technique, employed to evaluate the significant contributors influencing the output of any algorithm, highlights the distinctive aspect of the MDI as an exclusive feature within the realm of the RF algorithm.

Arsenic mechanism in groundwater

Arsenic levels in groundwater in the Vehari district varied from low to high levels of enrichment, as

Table 7 Importance features using permutation feature for ANN showing the change of the accuracy after a variable is eliminated

Variable	Accuracy variation for ANN
All variables	0.80
Eliminated Na^+ (mg/L)	0.58
Eliminated HCO_3^- (mg/L)	0.64
Eliminated NO_3^- (mg/L)	0.45
Eliminated SO_4^{2-} (mg/L)	0.57
Eliminated Cl^- (mg/L)	0.68
Eliminated TDS (mg/L)	0.80
Eliminated Ca^{2+} (mg/L)	0.60
Eliminated Mg^{2+} (mg/L)	0.52
Eliminated K^+ (mg/L)	0.60
Eliminated Fe^{2+} (mg/L)	0.62
Eliminated F^- (mg/L)	0.59
Eliminated pH	0.80

depicted in Fig. 9. Results indicated that 43.5% of the samples in the Vehari district exceeded the WHO 2022 permissible limit of As (10 $\mu\text{g/L}$). The correlation between As and some essential parameters was drawn to investigate the As release mechanism. The correlations are presented in scatter diagrams in Fig. 10. The results from the present study area showed some trend of oxidative desorption with an increased evaporative concentration mechanism concluded based on alkaline pH (6.7–8.2), low iron, high bicarbonates, high sulfates, negative correlation of iron with arsenic, respectively, and significant positive correlation between $As-HCO_3^-$ and $As-SO_4^{2-}$, and slight positive correlation with pH in groundwater of Vehari. The Gibbs diagram also justified the evaporative mechanism, which showed that evaporation is also a dominant natural phenomenon in controlling the water chemistry of the study area (Fig. 5). Ion-exchange processes and the adsorption of ions in the study region significantly impacted the concentration of As. Previous research has indicated that Ca^{2+} can potentially interfere with As adsorption due to the effect of ion reactions on mineral surfaces (Xie et al., 2008). High competing ionic compositions can thus aid arsenic desorption (HCO_3^- , SO_4^{2-} , Na^+ , K^+ , Mg^+ , and Ca^{2+}). The findings were consistent with previous studies of (Brahman et al., 2016) and (Shahab et al., 2019) with a high As concentration in Sindh province, Pakistan. Moreover, variations in As enrichment in high-pH groundwater could be attributed to soil salinization and subsurface environmental conditions (Li et al., 2020b). The weak correlation between As and pH observed in this study may be due to alkaline desorption, which can impact the release of As into groundwater. Additionally, the aquifers in the study region have been reported to be alluvial, composed of silt, sand, and gravel, and have elevated As levels in the Punjab province (Shahab et al., 2019). Punjab province has a high evaporation rate, with 74–80 percent of groundwater being highly evaporated (Yu et al., 2015). In this study, no statistically significant correlation between As, NO_3^- , and F^- was found, as their concentration levels were very low in almost all of the groundwater samples. The SI estimation facilitates understanding the reaction pathways and the measurement of mineral dissolution and precipitation. In the geochemical simulation model (Fig. 11), aquifer conditions were undersaturated ($SI < 0$) with calcium carbonate and rock salt

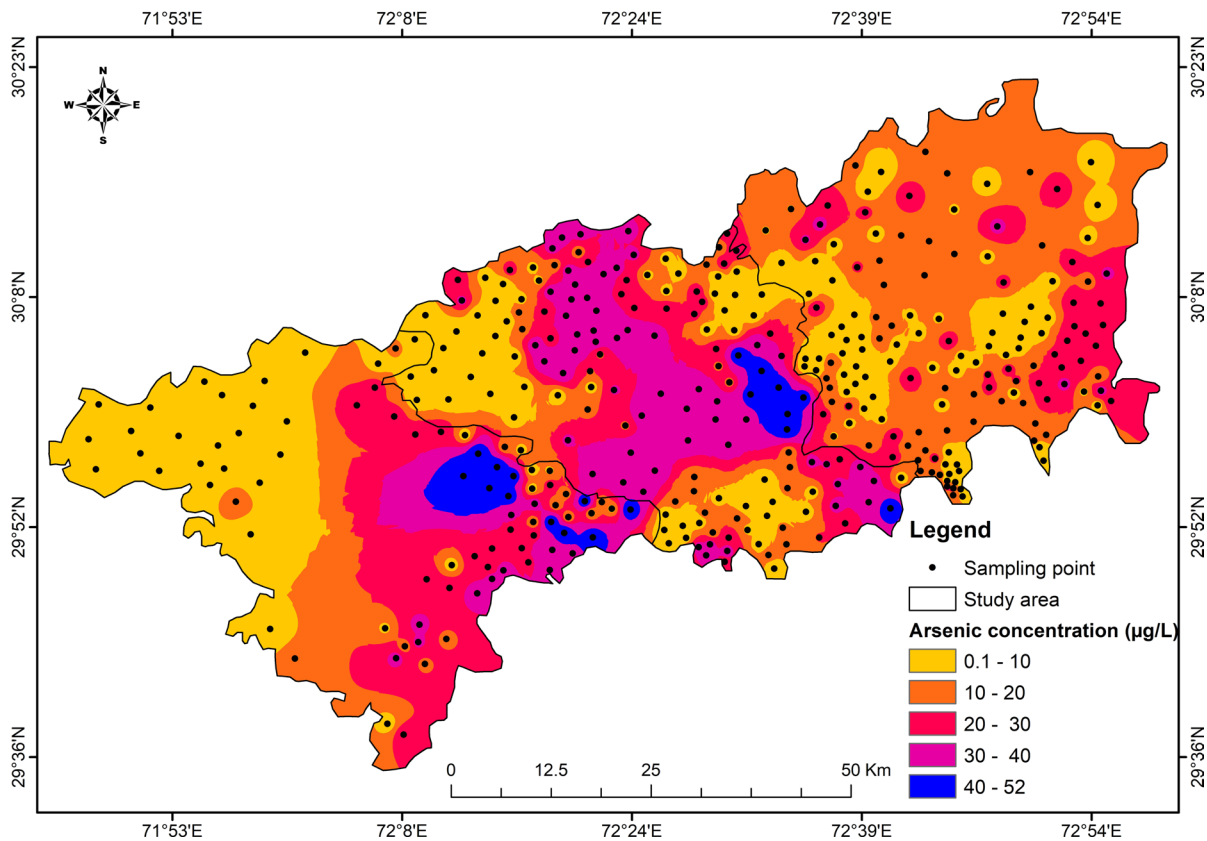


Fig. 9 Spatial distribution of groundwater As in study area

minerals, including aragonite, calcite, dolomite, and halite. These mineral phases had negative SI values and were unlikely to precipitate, but they may have played an important role in releasing As into aquifers through dissolution (Rashid et al., 2022). In contrast, the SI was positive for anhydrite, gypsum, and iron oxide mineral phases, including goethite and hematite. These minerals tended to participate in groundwater (Fig. 11). (Bhattacharya et al., 2009) found that iron oxides in the sediments of the flood plain in Bangladesh inhibited As mobility in groundwater.

WQI

The WQI is a popular method for determining groundwater quality for drinking (Narsimha & Sudarshan, 2017). The WQI was used to check the suitability of groundwater in the research region. The WQI is divided into five classes: excellent (50), good (50–100), poor (100–200), and unsuitable (>200).

Table 8 shows that the samples ($n=161$) were classified as "Poor" with 40.25 and 30% unsuitable contributions, while the samples 1.5 and 27.75% were classified as "Excellent" and "Good," respectively. Most samples had poor to unsuitable drinking water quality, showing that the study areas' groundwater sources are unsafe to drink. The water quality suitability map is depicted in Fig. 12.

Conclusions

The presence of high levels of As in drinking water sources can make it unsuitable for consumption. In the current study, 174 of the 400 samples (43.5%) had As concentrations that exceeded the permissible limit of 10 µg/L set by the World Health Organization (WHO, 2022) for drinking water. The As levels measured ranged from 0.1 to 52 µg/L. Ion-exchange processes and the adsorption of ions in the study

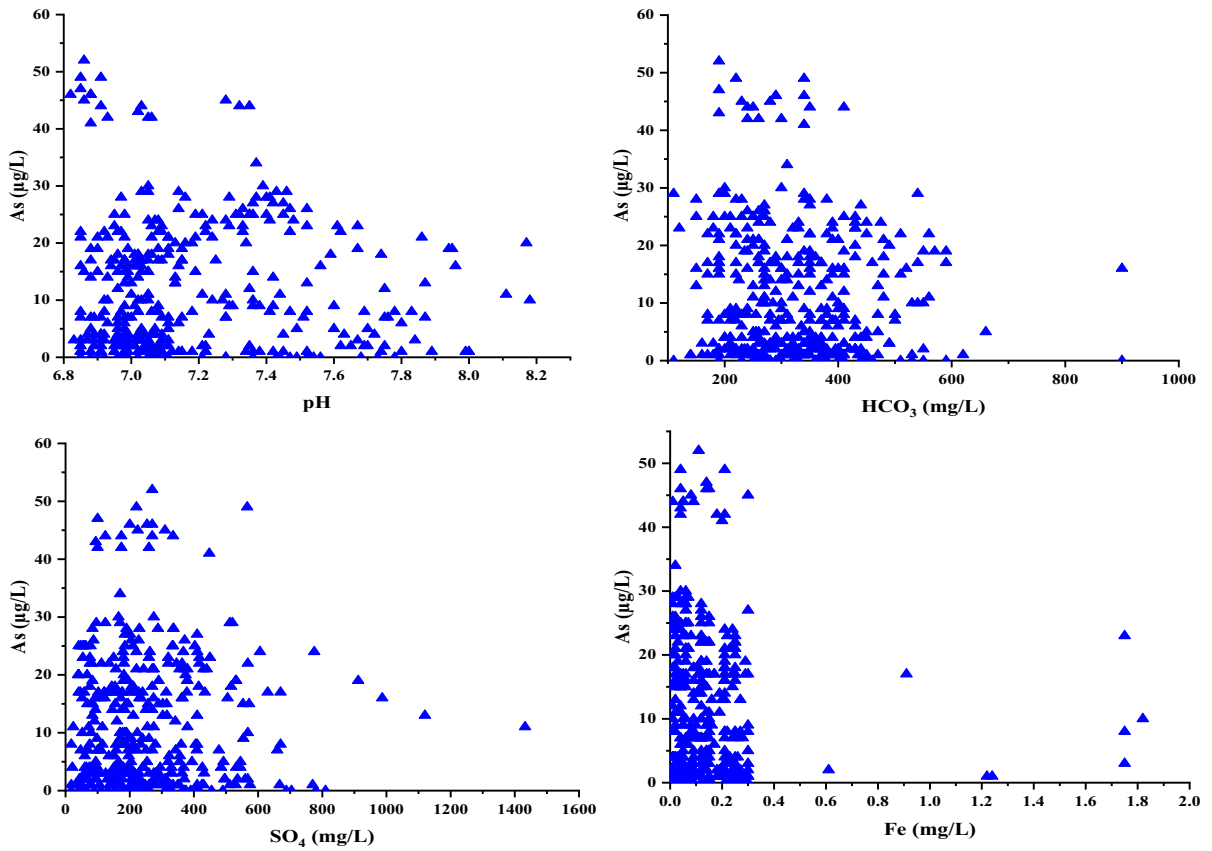


Fig. 10 Scatter diagram showing the correlation between arsenic and different variables in groundwater

Fig. 11 Relationships between As and saturation indices

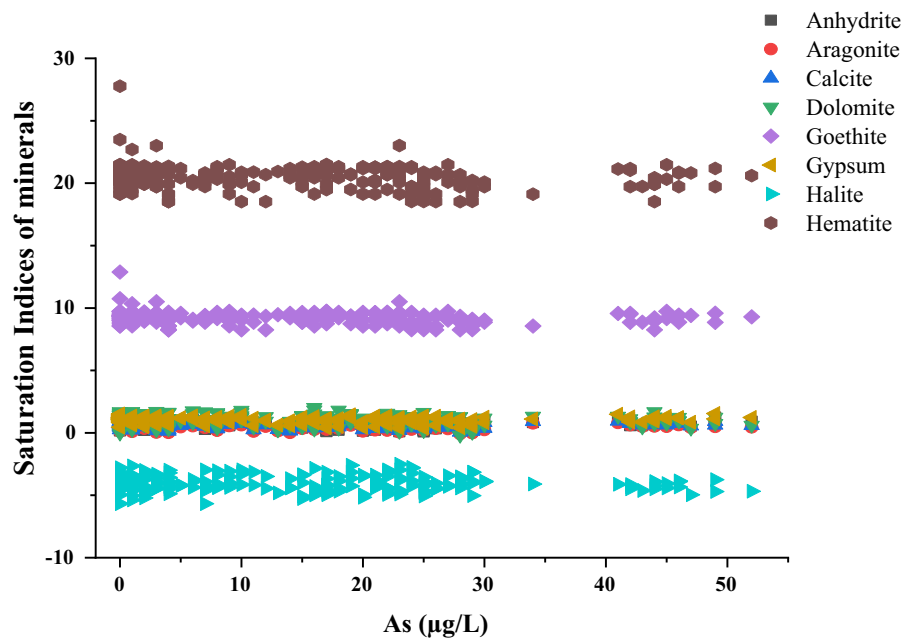
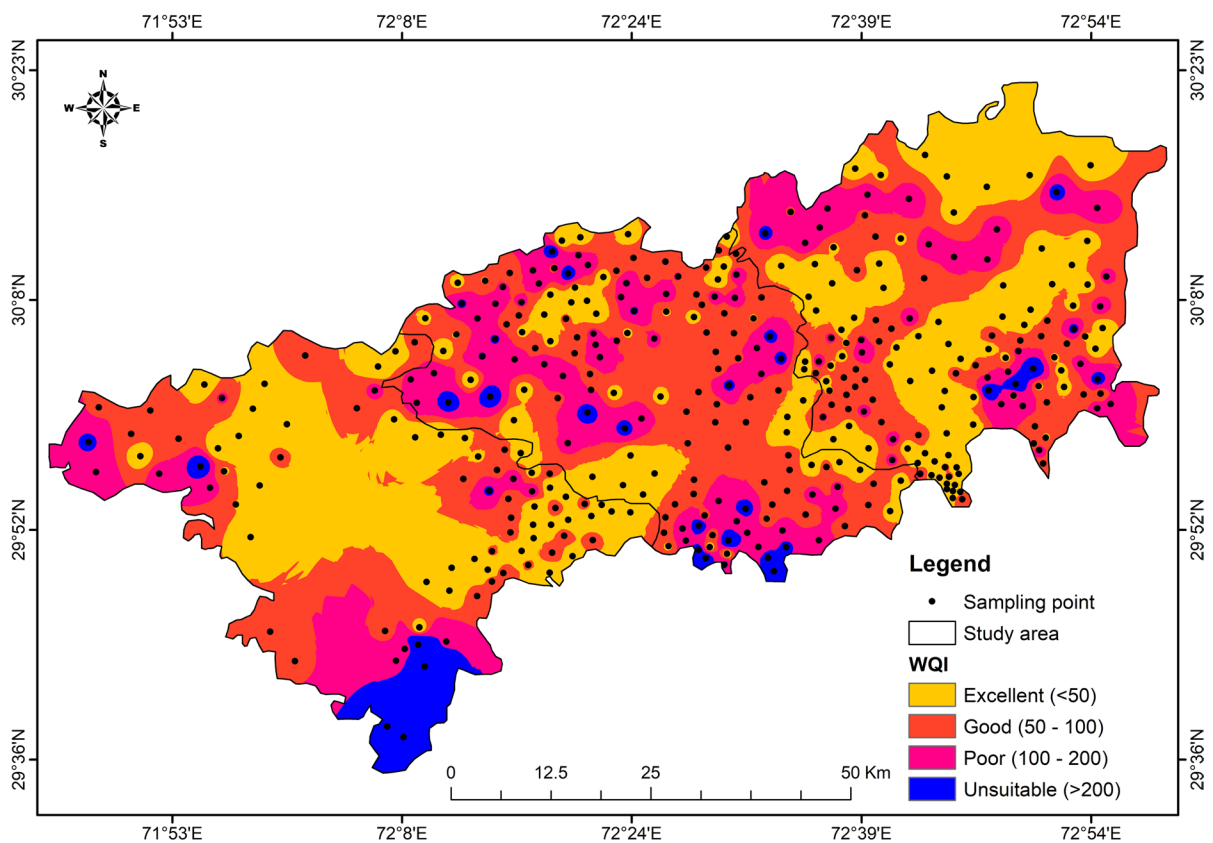


Table 8 Water Quality Index (WQI) classification of the study area

WQI five categories	No of samples	% of samples
Excellent (<50)	6	1.5
Good (>50),	111	27.75
Poor (>100)	161	40.25
Unsuitable (>200)	122	30.5

region significantly impacted the concentration of As. The elevated concentrations of basic physiochemical parameters, such as EC, TDS, HCO_3^- , and Na^+ , exceeded the permissible limits set by WHO, thereby rendering the water unsafe for drinking. Multivariate statistical approaches in the study suggest that geogenic and anthropogenic activities in the region cause As enrichment in groundwater. The hydrochemical analysis of groundwater samples indicates a combination of CaMgCl and CaCl types. The Gibbs plot

demonstrated that the prevailing rock composition substantially influences the groundwater's chemical makeup. Moreover, the results from geochemical modeling displayed that As had negative saturation indices with calcium carbonate and salt minerals, including aragonite, calcite, dolomite, and halite. According to the WQI, most of the water samples from the Vehari district had poor water quality. Artificial Neural Networks, Random Forest, and Logistic Regression machine learning techniques were used to predict As levels in the study region. Results indicate that the Random Forest technique was the most effective, with an accuracy of 0.85. The permutation feature and the MDI were employed to identify the variables influencing arsenic levels in the region. These approaches identified variables such as F^- , Fe^{2+} , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , SO_4^{2-} , NO_3^- , HCO_3^- , and Na^+ as contributing factors to As concentration. These findings suggest that the Random Forest model can be used as a reliable algorithm for forecasting

**Fig. 12** Groundwater suitability assessment for drinking purposes in the study area

groundwater arsenic in the Vehari region and can be extended to other locations for predicting groundwater contamination. However, future research should focus on developing more adaptive models to improve the accuracy of groundwater pollution prediction.

Acknowledgements The authors gratefully acknowledge the help of anonymous peer reviewers and the editors for their valuable comments.

Author Contributions JI: wrote the main manuscript. CS: Supervision. MA: data analysis. AR: conceptualization. ZU: made Fig. 1. HA: made Fig. 2. AN and MY, JB material preparation. AA and AU: reviewed the article.

Funding The research work was supported by the National Key R&D Program of China (2022YFC2503001) and the National Natural Science Foundation of China (Grant 42177078).

Data Availability Material The data will be provided on request to the corresponding author.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable.

Consent to participate All authors reviewed and approved the final manuscript.

Consent to publish All authors approved this for publication.

References

Abbas, F., Zhang, F., Ismail, M., Khan, G., Iqbal, J., Alrefaei, A. F., & Albeshr, M. F. (2023). Optimizing machine learning algorithms for landslide susceptibility mapping along the Karakoram Highway, Gilgit Baltistan, Pakistan: A comparative study of baseline, bayesian, and metaheuristic hyperparameter optimization techniques. *Sensors*, 23(15), 6843.

Abrahart, R.J., Kneale, P.E., Linda, M. (2005). *See: Neural Networks for Hydrological Modelling*. AA Balkema Publishers (Leiden, The Netherlands).

Adimalla, N., Li, P., & Qian, H. (2018). Evaluation of groundwater quality, Peddavagu in Central Telangana (PCT), South India: An insight of controlling factors of fluoride enrichment. *Modeling Earth Systems and Environment*, 4, 841–852.

Adimalla, N., Qian, H., & Tiwari, D. M. (2021). Groundwater chemistry, distribution and potential health risk appraisal of nitrate enriched groundwater: A case study from the semi-urban region of South India. *Ecotoxicology and Environmental Safety*, 207, 111277.

Afzaal, H., Farooque, A. A., Abbas, F., Acharya, B., & Esau, T. (2019). Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. *Water*, 12(1), 5.

Agbasi, J. C., & Egbueri, J. C. (2023). Intelligent soft computational models integrated for the prediction of potentially toxic elements and groundwater quality indicators: A case study. *Journal of Sedimentary Environments*, 8(1), 57–79.

Ahmad, N., Ahmad, M., Rafiq, M., Iqbal, N., Ali, M., & Sajjad, M. I. (2002). Hydrological modeling of the Lahore aquifer using isotopic chemical and numerical techniques. *Science Vision*, 7(3–4), 169–194.

Ahmadi, M., Motlagh, H. R., Jaafarzadeh, N., Mostoufi, A., Saeedi, R., Barzegar, G., & Jorfi, S. (2017). Enhanced photocatalytic degradation of tetracycline and real pharmaceutical wastewater using MWCNT/TiO2 nano-composite. *Journal of Environmental Management*, 186, 55–63.

Ali, A., Zahid, U., Maria, S., Junaid, G., Abdur, R., Warda, K., Muhammad, I., Ullah, K., & Waqas, A. (2023). Geochemical investigation of OCPs in the rivers along with drains and groundwater sources of Eastern Punjab, Pakistan. *Exposure and Health*, pp 1–16.

Al-Mukhtar, M. (2019). Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. *Environmental Monitoring and Assessment*, 191(11), 1–12.

Awu, J., Ogunjirin, O., Willoughby, F., & Adewumi, A. (2015). Potability evaluation of selected river waters in Ebonyi State, Nigeria. *Nigerian Journal of Technological Development*, 12(1), 27–35.

Azis, A. (2015). Conceptions and practices of assessment: A case of teachers representing improvement conception. *Teflin Journal*, 26(2), 129–154.

Baloch, M.Y.J., Su, C., Talpur, S.A., Iqbal, J., & Bajwa, K.J.J.o.E.S. (2022). Arsenic removal from groundwater using iron pyrite: Influences factors and removal mechanism. *Journal of Earth Science*, 6.

Bhattacharya, S., & Mishra, S. (2018). Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA. *Journal of Petroleum Science and Engineering*, 170, 1005–1017.

Bhattacharya, P., Hasan, M. A., Sracek, O., Smith, E., Ahmed, K. M., Von Brömssen, M., Imamul Huq, S. M. & Naidu, R. (2009). Groundwater chemistry and arsenic mobilization in the Holocene flood plains in south-central Bangladesh. *Environmental Geochemistry and Health*, 31, 23–43.

Boateng, T. K., Opoku, F., Acquah, S. O., & Akoto, O. (2016). Groundwater quality assessment using statistical approach and water quality index in Ejisu-Juaben Municipality, Ghana. *Environmental Earth Sciences*, 75(6), 1–14.

Brahman, K. D., Kazi, T. G., Afridi, H. I., Arain, S. S., Kazi, A. G., Talpur, F. N., Baig, J. A., Panhwar, A. H., Arain, M. S., Ali, J., Arain, M. B. & Naemullah. (2016). Toxic risk assessment of arsenic in males through drinking water in Tharparkar Region of Sindh, Pakistan. *Biological Trace Element Research*, 172(1), 61–71.

- Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48(1), 287–297.
- Cao, H., Xie, X., Wang, Y., & Liu, H. (2022). Predicting geogenic groundwater fluoride contamination throughout China. *Journal of Environmental Sciences*, 115, 140–148.
- Chae, Y. T., Horesh, R., Hwang, Y., & Lee, Y. M. (2016). Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings*, 111, 184–194.
- Chidambaram, S., Sarathidasan, J., Srinivasamoorthy, K., Thivya, C., Thilagavathi, R., Prasanna, M. V., Singaraja, C., & Nepolian, M. (2018). Assessment of hydrogeochemical status of groundwater in a coastal region of Southeast coast of India. *Applied Water Science*, 8(1), 1–14.
- Çiner, F., Sunkari, E. D., & Şenbaş, B. A. (2021). Geochemical and multivariate statistical evaluation of trace elements in groundwater of Niğde Municipality, South-Central Turkey: Implications for arsenic contamination and human health risks assessment. *Archives of Environmental Contamination and Toxicology*, 80(1), 164–182.
- Coulibaly, P., Ancil, F., & Bobée, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology*, 230(3–4), 244–257.
- Dilpazeer, F., Munir, M., Baloch, M. Y. J., Shafiq, I., Iqbal, J., Saeed, M., Abbas, M. M., Shafique, S., Aziz, K. H. H., Mustafa, A., & Mahboob, I. (2023). A comprehensive review of the latest advancements in controlling arsenic contaminants in groundwater. *Water*, 15(3), 478.
- Egbueri, J. C. (2021). Prediction modeling of potentially toxic elements' hydrogeopollution using an integrated Q-mode HCs and ANNs machine learning approach in SE Nigeria. *Environmental Science and Pollution Research*, 28(30), 40938–40956.
- Egbueri, J. C., & Agbasi, J. C. (2022a). Combining data-intelligent algorithms for the assessment and predictive modeling of groundwater resources quality in parts of southeastern Nigeria. *Environmental Science and Pollution Research*, 29(38), 57147–57171.
- Egbueri, J. C., & Agbasi, J. C. (2022b). Data-driven soft computing modeling of groundwater quality parameters in southeast Nigeria: Comparing the performances of different algorithms. *Environmental Science and Pollution Research*, 29(25), 38346–38373.
- El Bilali, A., Taleb, A., & Brouziyne, Y. (2021). Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricultural Water Management*, 245, 106625.
- Erguzel, T. T., Noyan, C. O., Eryilmaz, G., Ünsalver, B. Ö., Cebi, M., Tas, C., Dilbaz, N., & Tarhan, N. (2019). Binomial logistic regression and artificial neural network methods to classify opioid-dependent subjects and control group using quantitative EEG power measures. *Clinical EEG and Neuroscience*, 50(5), 303–310.
- Fatima, S., Hussain, I., Rasool, A., Xiao, T., & Farooqi, A. (2018). Comparison of two alluvial aquifers shows the probable role of river sediments on the release of arsenic in the groundwater of district Vehari, Punjab, Pakistan. *Environmental Earth Sciences*, 77, 1–14.
- Fornes, O., et al. (2020). JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1), D87–D92.
- Ghani, J., et al. (2022). Hydrogeochemical characterization, and suitability assessment of drinking groundwater: Application of geostatistical approach and geographic information system. *Frontiers in Environmental Science*, 10, 874464.
- Gopinath, S., et al. (2018). Hydrochemical characteristics and salinity of groundwater in parts of Nagapattinam district of Tamil Nadu and the Union Territory of Puducherry, India. *Carbonates and Evaporites*, 33(1), 1–13.
- He, S., Wu, J., Wang, D., & He, X. (2022). Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere*, 290, 133388.
- Herczeg, A., Dogramaci, S., & Leaney, F. (2001). Origin of dissolved salts in a large, semi-arid groundwater system: Murray Basin, Australia. *Marine and Freshwater Research*, 52(1), 41–52.
- Hussain, M. A., Chen, Z., Wang, R., & Shoaib, M. (2021). PS-InSAR-based validated landslide susceptibility mapping along Karakorum Highway, Pakistan. *Remote Sensing*, 13(20), 4129.
- Hussain, M. A., et al. (2022). Landslide susceptibility mapping using machine learning algorithm validated by persistent scatterer In-SAR technique. *Sensors*, 22(9), 3119.
- Iqbal, J., Ali, M., Ali, A., Raza, D., Bashir, F., Ali, F., Hussain, S., & Afzal, Z. (2020). Investigation of cryosphere dynamics variations in the upper indus basin using remote sensing and gis. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44, 59–63.
- Iqbal, J., Amin, G., Su, C., Haroon, E., & Jat Baloch, M.Y. (2023a). Assessment of landcover impacts on the groundwater quality using hydrogeochemical and geospatial techniques. *Environmental Science and Pollution-Research International*.
- Iqbal, J. et al. (2023b). Groundwater fluoride and nitrate contamination and associated human health risk assessment in South Punjab, Pakistan. *Environmental Science and Pollution Research*, 30(22), 61606–61625.
- Jamil, A. Khan, A. A., Bayram, B., Iqbal, J., Amin, G., Yesiltepe, M., & Hussain, D. (2019). Spatio-temporal glacier change detection using deep learning: a case study of Shishper Glacier in Hunza. In: *International Symposium on Applied Geoinformatics*, 5.
- Jat Baloch, M. Y., & Mangi, S. H. (2019). Treatment of synthetic greywater by using banana, orange and sapodilla peels as a low cost activated carbon. *Journal of Material and Environmental Science*, 10(10), 966–986.
- Jat Baloch, M. Y., Talpur, S. A., Talpur, H. A., Iqbal, J., Mangi, S. H., & Memon, S. (2020). Effects of arsenic toxicity on the environment and its remediation techniques: A review. *Journal of Water and Environment Technology*, 18(5), 275–289.
- Jat Baloch, M. Y., Zhang, W., Chai, J., Li, S., Alqurashi, M., Rehman, G., Tariq, A., Talpur, S. A., Iqbal, J., Munir, M., & Hussein, E. E. (2021). Shallow groundwater

- quality assessment and its suitability analysis for drinking and irrigation purposes. *Water*, 13(23), 3361.
- Jat Baloch, M. Y., Su, C., Talpur, S. A., Iqbal, J., & Bajwa, K. (2023). Arsenic removal from groundwater using iron pyrite: Influence factors and removal mechanism. *Journal of Earth Science*, pp. 1–11.
- Jat Baloch, M. Y., Zhang, W., Al Shoumik, B. A., Nigar, A., Elhassan, A. A., Elshekh, A. E., & Iqbal, J. (2022a). Hydrogeochemical mechanism associated with land use land cover indices using geospatial, remote sensing techniques, and health risks model. *Sustainability*, 14(24), 16768.
- Jat Baloch, M. Y., Zhang, W., Zhang, D., Al Shoumik, B. A., Iqbal, J., Li, S., Chai, J., Farooq, M. A., & Parkash, A. (2022b). Evolution mechanism of arsenic enrichment in groundwater and associated health risks in southern Punjab, Pakistan. *International Journal of Environmental Research and Public Health*, 19(20), 13325.
- Jia, C., Altaf, A. R., Li, F., Ashraf, I., Zafar, Z., & Nadeem, A. A. (2023). Comprehensive assessment on groundwater quality, pollution characteristics, and ecological health risks under seasonal thaws: Spatial insights with Monte Carlo simulations. *Groundwater for Sustainable Development*, 22, 100952.
- Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering Trends and Technology*, 3(6), 714–717.
- Khalid, S., Shahid, M., Natasha, Shah, A. H., Saeed, F., Ali, M., Qaisrani, S. A., & Dumat, C. (2020). Heavy metal contamination and exposure risk assessment via drinking groundwater in Vehari, Pakistan. *Environmental Science and Pollution Research*, 27, 39852–39864.
- Khan, N., Bano, A., & Zandi, P. (2018). Effects of exogenously applied plant growth regulators in combination with PGPR on the physiology and root growth of chickpea (*Cicer arietinum*) and their role in drought tolerance. *Journal of Plant Interactions*, 13(1), 239–247.
- Kim, H.-Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2), 152–155.
- Li, C., Sanchez, G. M., Wu, Z., Cheng, J., Zhang, S., Wang, Q., Li, F., Sun, G., & Meentemeyer, R. K. (2020a). Spatiotemporal patterns and drivers of soil contamination with heavy metals during an intensive urbanization period (1989–2018) in southern China. *Environmental Pollution*, 260, 114075.
- Li, Z., Yang, Q., Yang, Y., Xie, C., & Ma, H. (2020b). Hydrogeochemical controls on arsenic contamination potential and health threat in an intensive agricultural area, northern China. *Environmental Pollution*, 256, 113455.
- Li, S., Zhang, W., Zhang, D., Xiu, W., Wu, S., Chai, J., Ma, J., Jat Baloch, M. F., Sun, S., & Yang, Y. (2023). Migration risk of *Escherichia coli* O157: H7 in unsaturated porous media in response to different colloid types and compositions. *Environmental Pollution*, 323, 121282.
- Mallick, J., Kumar, A., Almesfer, M. K., Alsubih, M., Singh, C. K., Ahmed, M., & Khan, R. A. (2021). An index-based approach to assess groundwater quality for drinking and irrigation in Asir region of Saudi Arabia. *Arabian Journal of Geosciences*, 14(3), 1–17.
- Manjusree, T., Joseph, S., & Thomas, J. (2009). Hydrogeochemistry and groundwater quality in the coastal sandy clay aquifers of Alappuzha district, Kerala. *Journal of the Geological Society of India*, 74(4), 459–468.
- Markwart, R., Willrich, N., Haller, S., Noll, I., Koppe, U., Werner, G., Eckmanns, T., & Reuss, A. (2019). The rise in vancomycin-resistant *Enterococcus faecium* in Germany: Data from the German Antimicrobial Resistance Surveillance (ARS). *Antimicrobial Resistance & Infection Control*, 8(1), 1–11.
- Matayoshi, J., Uzun, H., & Cosyn, E. (2019). Deep (un) learning: Using neural networks to model retention and forgetting in an adaptive learning system. In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 258–269.
- Mitchell, P., Liew, G., Gopinath, B., & Wong, T. Y. (2018). Age-related macular degeneration. *The Lancet*, 392(10153), 1147–1159.
- Musa, R. M., Majeed, A. A., Taha, Z., Abdullah, M. R., Maliki, A. H. M., & Kosni, N. A. (2019). The application of Artificial Neural Network and k-Nearest Neighbour classification models in the scouting of high-performance archers from a selected fitness and motor skill performance parameters. *Science & Sports*, 34(4), e241–e249.
- Nafouanti, M. B., Li, J., Mustapha, N. A., Uwamungu, P., & Dalal, A.-A. (2021). Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: Comparison of random forest, logistic regression and artificial neural network. *Applied Geochemistry*, 132, 105054.
- Narsimha, A., & Sudarshan, V. (2017). Contamination of fluoride in groundwater and its effect on human health: A case study in hard rock aquifers of Siddipet, Telangana State, India. *Applied Water Science*, 7(5), 2501–2512.
- Omeka, M. E., & Egbueri, J. C. (2023). Hydrogeochemical assessment and health-related risks due to toxic element ingestion and dermal contact within the Nnewi-Awka urban areas, Nigeria. *Environmental Geochemistry and Health*, 45(5), 2183–2211.
- Onyemesili, O. O., Egbueri, J. C., & Ezugwu, C. K. (2022). Assessing the pollution status, ecological and health risks of surface waters in Nnewi urban, Nigeria: Implications of poor waste disposal. *Environmental Forensics*, 23(3–4), 346–360.
- Otchere, D. A., Ganat, T. O. A., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal of Petroleum Science and Engineering*, 200, 108182.
- Rahman, A., Mondal, N., & Fauzia, F. (2021). Arsenic enrichment and its natural background in groundwater at the proximity of active floodplains of Ganga River, northern India. *Chemosphere*, 265, 129096.
- Raju, N. J. J. T. R. (2006). Iron contamination in groundwater: A case from Tirumala-Tirupati environs. *India*, 1(1), 28–31.
- Ranjgar, B., Razavi-Termeh, S. V., Foroughnia, F., Sadeghi-Niaraki, A., & Perissin, D. (2021). Land subsidence susceptibility mapping using persistent scatterer sar interferometry technique and optimized hybrid machine learning algorithms. *Remote Sensing*, 13(7), 1326.

- Rashid, A., Guan, D. X., Farooqi, A., Khan, S., Zahir, S., Jehan, S., Khattak, S. A., Khan, M. S., & Khan, R. I. (2018a). Fluoride prevalence in groundwater around a fluorite mining area in the flood plain of the River Swat, Pakistan. *Science of the Total Environment*, *635*, 203–215.
- Rashid, A., Guan, D.-X., Farooqi, A., Khan, S., Zahir, S., Jehan, S., Anjum Khattak, S., Sufaid Khan, M., Khan, R. (2018b). Fluoride prevalence in groundwater around a fluorite mining area in the flood plain of the River Swat, Pakistan. *Science of the Total Environment*, *635*, 203–215.
- Rashid, A., Khattak, S. A., Ali, L., Zaib, M., Jehan, S., Ayub, M., & Ullah, S. (2019). Geochemical profile and source identification of surface and groundwater pollution of District Chitral, Northern Pakistan. *Microchemical Journal*, *145*, 1058–1065.
- Rashid, A., Farooqi, A., Gao, X., Zahir, S., Noor, S., & Khattak, J. A. (2020). Geochemical modeling, source apportionment, health risk exposure and control of higher fluoride in groundwater of sub-district Dargai, Pakistan. *Chemosphere*, *243*, 125409.
- Rashid, A., Ayub, M., Khan, S., Ullah, Z., Ali, L., Gao, X., Li, C., El-Serehy, H. A., Kaushik, P. & Rasool, A. (2022). Hydrogeochemical assessment of carcinogenic and non-carcinogenic health risks of potentially toxic elements in aquifers of the Hindukush ranges, Pakistan: Insights from groundwater pollution indexing, GIS-based, and multivariate statistical approaches. *Environmental Science and Pollution Research*, *29*(50), 75744–75768.
- Rashid, A., Ayub, M., Ullah, Z., Ali, A., Sardar, T., Iqbal, J., Gao, X., Bundschuh, J., Li, C., Khattak, S. A., Ali, L., El-Serehy, H. A., Kaushik, P., & Khan, S. (2023). Groundwater Quality, Health Risk Assessment, and Source Distribution of Heavy Metals Contamination around Chromite Mines: Application of GIS, Sustainable Groundwater Management, Geostatistics, PCAMLR, and PMF Receptor Model. *International Journal of Environmental Research and Public Health*, *20*(3), 2113.
- Rehman, G., Ain, Q. T., Zaheer, M., Bao, L., & Iqbal, J. (2019). An analysis of two-dimensional flow through a water reservoir using mathematical approach. *Reviews of Environmental Contamination ECR*, *2*(1), 11–13.
- Rizeei, H. M., Azeez, O. S., Pradhan, B., & Khamees, H. H. (2018). Assessment of groundwater nitrate contamination hazard in a semi-arid region by using integrated parametric IPNOA and data-driven logistic regression models. *Environmental Monitoring and Assessment*, *190*(11), 1–17.
- Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment*, *476*, 189–206.
- Roy, S. K., Shekhar, V., Lassar, W. M., & Chen, T. (2018). Customer engagement behaviors: The role of service convenience, fairness and quality. *Journal of Retailing and Consumer Services*, *44*, 293–304.
- Sahin, R., Kumar, A., Chandrakar, R., Michalska-Domańska, M., & Dubey, V. (2021). 3 The physico-chemical interaction of fluorine with the environment. *Water Resource Technology: Management for Engineering Applications*, p 17.
- Salem, Z. E., Atwia, M. G., & El-Horiny, M. M. (2015). Hydrogeochemical analysis and evaluation of groundwater in the reclaimed small basin of Abu Mina, Egypt. *Hydrogeology Journal*, *23*(8), 1781–1797.
- Selvakumar, S., Chandrasekar, N., & Kumar, G. (2017). Hydrogeochemical characteristics and groundwater contamination in the rapid urban development areas of Coimbatore, India. *Water Resources and Industry*, *17*, 26–33.
- Shah, A. H., Shahid, M., Khalid, S., Natasha, Shabbir, Z., Bakhat, H. F., Murtaza, B., Farooq, A., Akram, M., Shah, G. M., Nasim, W., & Niazi, N. K. (2020). Assessment of arsenic exposure by drinking well water and associated carcinogenic risk in peri-urban areas of Vehari, Pakistan. *Environmental Geochemistry and Health*, *42*, 121–133.
- Shahab, A., Qi, S., & Zaheer, M. (2019). Arsenic contamination, subsequent water toxicity, and associated public health risks in the lower Indus plain, Sindh province, Pakistan. *Environmental Science and Pollution Research*, *26*(30), 30642–30662.
- Shahid, M., Khalid, M., Dumat, C., Khalid, S., Niazi, N. K., Imran, M., Bibi, I., Iftikhar, A., Hafiz, H., Tabassum, R. A. (2018a). Arsenic level and risk assessment of groundwater in Vehari, Punjab Province, Pakistan. *Exposure and Health*, *10*, 229–239.
- Shahid, M., et al. (2018b). A meta-analysis of the distribution, sources and health risks of arsenic-contaminated groundwater in Pakistan. *Environmental Pollution*, *242*, 307–319.
- Stojanović Bjelić, L., Ilić, P., Nešković Markić, D., Ilić, S., Popović, Z., Mrazovac Kurilić, S., Mihajlović, D., Farooqi, Z. U. R., Jat Baloch, M. Y., Mohamed, M. H & Ahmed, M. (2023). Contamination in water and ecological risk of heavy metals near a coal mine and a thermal power plant (Republic of Srpska, Bosnia and Herzegovina). *Applied Ecology and Environmental Research*, *21*(5), 3807–3822.
- Sun, A. Y., & Scanlon, B. R. (2019). How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters*, *14*(7), 073001.
- Tabassum, R. A., Shahid, M., Dumat, C., Niazi, N. K., Khalid, S., Shah, N. S., & Khalid, S. (2019). Health risk assessment of drinking arsenic-containing groundwater in Hasilpur, Pakistan: Effect of sampling area, depth, and source. *Environmental Science and Pollution Research*, *26*(20), 20018–20029.
- Tahir, M., & Rasheed, H. (2013). Fluoride in the drinking water of Pakistan and the possible risk of crippling fluorosis. *Drinking Water Engineering and Science*, *6*(1), 17–23.
- Tariq, A., Mumtaz, F., Zeng, X., Baloch, M. Y. J., & Moazzam, M. F. U. (2022). Spatio-temporal variation of seasonal heat islands mapping of Pakistan during 2000–2019, using day-time and night-time land surface temperatures MODIS and meteorological stations data. *Remote*

- Sensing Applications: Society and Environment*, 27, 100779.
- Tariq, A., Ali, S., Basit, I., Jamil, A., Farmonov, N., Khorrani, B., Khan, M. M., Sadri, S., Jat Baloch, M. Y., Islam, F., Junaid, M. B., & Hatamleh, W. A. (2023). Terrestrial and groundwater storage characteristics and their quantification in the Chitral (Pakistan) and Kabul (Afghanistan) river basins using GRACE/GRACE-FO satellite data. *Groundwater for Sustainable Development*, 23, 100990.
- Tropea, E., Hynds, P., McDermott, K., Brown, R.S., & Majury, A. (2021). Environmental adaptation of *E. coli* within private groundwater sources in southeastern Ontario: Implications for groundwater quality monitoring and human health. *Environmental Pollution*, 285, 117263.
- Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, 13(4), 397–405.
- Ullah, Z., Talib, M. A., Rashid, A., Ghani, J., Shahab, A., Irfan, M., Rauf, A., Bawazeer, S., Almarhoon, Z., & Mabkhot, Y. N. (2021). Hydrogeochemical investigation of elevated arsenic based on entropy modeling, in the aquifers of District Sanghar, Sindh, Pakistan. *Water*, 13(23), 3477.
- Ullah, Z., et al. (2022a). Groundwater contamination through potentially harmful metals and its implications in groundwater management. *Frontiers in Environmental Science*, 10, 2077.
- Ullah, Z., Xu, Y., Zeng, X. C., Rashid, A., Ali, A., Iqbal, J., Almutairi, M. H., Aleya, L., Abdel-Daim, M. M., & Shah, M. (2022b). Non-carcinogenic health risk evaluation of elevated fluoride in groundwater and its suitability assessment for drinking purposes based on water quality index. *International Journal of Environmental Research and Public Health*, 19(15), 9071.
- Ullah, Z., Rashid, A., Ghani, J., Talib, M. A., Shahab, A., & Lun, L. (2022c). Arsenic contamination, water toxicity, source apportionment, and potential health risk in groundwater of Jhelum Basin, Punjab, Pakistan. *Biological Trace Element Research*, 201(1), 514–524.
- Üstün, B., Melssen, W., Oudenhuijzen, M., & Buydens, L. (2005). Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta*, 544(1–2), 292–305.
- Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. *Psychometrika*, 61(3), 401–425.
- Wu, D., Wu, T., Liu, Q., & Yang, Z. (2020). The SARS-CoV-2 outbreak: What we know. *International Journal of Infectious Diseases*, 94, 44–48.
- WHO (2022). Guidelines for drinking-water quality: incorporating the first and second addenda.
- Xie, X., Wang, Y., Su, C., Liu, H., Duan, M., & Xie, Z. (2008). Arsenic mobilization in shallow aquifers of Datong Basin: Hydrochemical and mineralogical evidences. *Journal of Geochemical Exploration*, 98(3), 107–115.
- Xue-Jie, G., Mei-Li, W., & Giorgi, F. (2013). Climate change over China in the 21st century as simulated by BCC_CSM1.1-RegCM4.0. *Atmospheric and Oceanic Science Letters*, 6(5), 381–386.
- Yesilnacar, M. I., Sahinkaya, E., Naz, M., & Ozkaya, B. (2008). Neural network prediction of nitrate in groundwater of Harran Plain, Turkey. *Environmental Geology*, 56(1), 19–25.
- Yu, Q., Wang, Y., Xie, X., Currell, M., Pi, K., & Yu, M. (2015). Effects of short-term flooding on arsenic transport in groundwater system: A case study of the Datong Basin. *Journal of Geochemical Exploration*, 158, 1–9.
- Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(2), 56–70.
- Zhang, H., Cheng, S., Li, H., Fu, K., & Xu, Y. (2020). Groundwater pollution source identification and apportionment using PMF and PCA-APCA-MLR receptor models in a typical mixed land-use area in Southwestern China. *Science of the Total Environment*, 741, 140383.
- Zhang, W., Zhu, Y., Gu, R., Liang, Z., Xu, W., & Jat Baloch, M. Y. (2022). Health Risk Assessment during In Situ Remediation of Cr (VI)-Contaminated Groundwater by Permeable Reactive Barriers: A Field-Scale Study. *International Journal of Environmental Research Public Health*, 19(20), 13079.
- Zhang, W., Chai, J., Li, S., Wang, X., Wu, S., Liang, Z., Jat Baloch, M. Y., Silva, L. F. O., & Zhang, D (2022). Physiological characteristics, geochemical properties and hydrological variables influencing pathogen migration in subsurface system: What we know or not? *Geoscience Frontiers*, 13(6), 101346.
- Zhou, Y., Li, P., Chen, M., Dong, Z., & Lu, C. (2021). Groundwater quality for potable and irrigation uses and associated health risk in southern part of Gu'an County, North China Plain. *Environmental Geochemistry and Health*, 43(2), 813–835.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.