Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning-Based Assessment of Geochemical Controls on Groundwater Arsenic Hazard in Bangladesh

Tasnia Ashrafi Heya
Sinclair Community College, United States of America
Tahsin Habib
Texas A&M University
Shubhra Bhattacharjee
Texas Tech University
Md Shakil Kashem
Kansas State University
Tanvir Mustafy
North South University
Asef Mohammad Redwan
asef.redwan@northsouth.edu

North South University

Article

Keywords: arsenic hazard, groundwater quality, biogeochemical parameters, machine learning

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11548495/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning-Based Assessment of Geochemical Controls on Groundwater Arsenic Hazard in Bangladesh

Tasnia Ashrafi Heya[1], Tahsin Habib[2,3*], Shubhra Bhattacharjee[4], Md Shakil Kashem[5], Tanvir Mustafy[2] and Asef Mohammad Redwan[2**]*
Sinclair Community College, Dayton, Ohio, United States of America, 45402[1]
North South University, Dhaka, Bangladesh, 1229[2]
Texas A&M University, College Station, Texas, United States of America, 77840[3]
Texas Tech University, Lubbock, Texas, United States of America, 79409[4]
Kansas State University, Manhattan, Kansan, United States of America, 66506[5]
[*]authors contributed equally; [**]Corresponding author: asef.redwan@northsouth.edu

Abstract: Arsenic contamination in groundwater is a major global public health concern, affecting millions of people, particularly in Bangladesh. Understanding the mechanisms controlling arsenic mobilization is therefore essential for hazard assessment and mitigation. This study presents assessment of groundwater arsenic hazard by examining its association with selected geochemical factors through statistical analysis, machine learning, and spatial interpolation techniques. Although previous studies have applied machine learning to groundwater arsenic prediction in Bangladesh, the combined evaluation of geochemical associations, model performance, and spatial validation remains limited. To address this gap, principal component analysis (PCA), ANOVA F-test feature selection, Chi-square scoring function, and L-index were used to identify geochemical variables associated with arsenic occurrence. Three supervised machine learning algorithms were then applied to classify arsenic hazard levels. Among the individual predictors, Fe provided the highest maximum accuracy for As-level classification at approximately 92%, while P and Mn achieved accuracies of approximately 89%; sulfate also demonstrated predictive capability, although its spatial relationship with As was less consistent. The model outputs were further compared with spatial arsenic hotspots derived from Kriging interpolation. The findings improve understanding of the geochemical factors associated with arsenic mobilization and demonstrate the usefulness of data-driven models for groundwater arsenic hazard assessment.

Keywords: arsenic hazard, groundwater quality, biogeochemical parameters, machine learning

1. Introduction

Arsenic (As) is a carcinogenic and toxic metalloid found in sediments, soils, and groundwater.[1] Groundwater enriched with As from geogenic sources poses a risk of arsenic exposure to more than 230 million people worldwide. Bangladesh has a standard value of As concentration in drinking water five times (50 μg/L) higher than the WHO (World Health Organization) recommendation (10 μg/L).[2] Bangladesh has experienced one of the largest documented poisoning of a population exposures to groundwater arsenic contamination.[3] Approximately 35 to 75 million people in 54 out of 65 districts (comprising 126,134 km^2) are still exposed to severe arsenic levels (> 50 μg/L) in Bangladesh due to naturally occurring arsenic released from arsenic-bearing aquifer sediments.[4,5,6,3] Bangladesh has an arsenic concentration ranging between 3 and 315 μg/L in its alluvial aquifer; however, some areas have observed concentrations up to 2000 μg/L.[7,8] Public health risk remains particularly severe in rural areas, where many communities rely on

hand-pumped groundwater from shallow aquifers for drinking and domestic use.[6,5] Ingestion of As contaminated water is linked with skin cancer, cardiovascular disease, arsenicosis, diabetes, anemia, and neurological and reproductive disorders.[9] A national survey reported more than 40,000 arsenicosis patients in Bangladesh.[10] However, public health experts estimated that more than 2.5 million cases would be reported in the next 50 years because of arsenic in drinking water.[11]

The source of arsenic in Bangladesh is mainly geogenic, originating from arsenic-bearing minerals and sediments, including sulfide minerals such as pyrite (FeS_2) and arsenopyrite (FeAsS).[12] However, human activities, including excessive groundwater extraction, use of fertilizers and pesticides in agricultural lands, coal combustion, industrial and mining operations, can intensify As concentration in groundwater.[13] Sediment deposition by the Ganges, Brahmaputra, and Meghna (GBM) influences the spatial distribution of arsenic in soil and groundwater.[14,15] As concentrations are generally higher in central, southern, and southeastern parts of Bangladesh, northwestern and uplifted areas in the north-central region of the country are least affected (Reference: BGS and DPHE). Aquifer hydrogeochemical properties, the effect of oxidized and reduced mineral phases, and the presence of co-factors that interact with arsenic-rich solid phases can influence the mobilization of arsenic from sediments in groundwater.[16] Therefore, many researchers attempted to identify the detailed association between arsenic distribution and geochemical factors on a large scale in Bangladesh.[14,17,18,7] According to Nickson et al., iron (Fe) and manganese (Mn) rich minerals within alluvial sediment release arsenic in groundwater by the dissolution and desorption from Fe/Mn oxides through microbial action under reducing conditions.[19] Smedley et al. also reported that As desorption can be accelerated in the presence of high levels of phosphate, bicarbonate, silicate, and possibly organic matter because these species can compete with arsenic for mineral adsorption sites.[1] Throughout the GBM delta, groundwater flow is slow, and the sedimentary formations of relatively young aquifers are rich in natural organic matter (NOM).[12] The presence of NOM could outcompete phosphate (PO_4^{3-}) toward sorption sites on soil constituents and decrease phosphate (PO_4^{3-}) sorption, thus releasing phosphorus (P) during the oxidation procedure.[13] Smith et al. found that the increment of P (0.16 to 3.2 mmol/L) concentration substantially reduced As (0.63 to 0.37 mmol/Kg) sorption in the soil, and therefore; increased As concentration in the groundwater.[20] Hence, NOM plays a vital role in arsenic mobilization in the aquifer; however, the direct relationship between PO_4^{3-} and arsenic concentration is not well understood.

The sulfur cycle may also play an important but complex role in arsenic mobilization. A study of Canfield et al., (2010) identified that sulfur cycling in anoxic conditions influences the biogeochemical process by simultaneously reducing sulfate and oxidating sulfide.[21] Sulfate-reducing bacteria: *Desulfovibrio desulfuricans,* and *Desulfovibrio vulgaris* were detected in many areas in Bangladesh, which influences the As release in the shallow groundwater from sulfur-bearing minerals.[18] The result of the study of Guo et

al. supported this hypothesis and identified an association between cryptic sulfur cycling and arsenic release in groundwater.[22] They identified that there are two crucial steps of sulfur cycling: sulfide produced by bacterial sulfate reduction (BSR) and reductive dissolution of Fe (III) oxyhydroxides through bacterial disproportionation of sulfur intermediates increase groundwater arsenic concentration.[22] Moreover, in groundwater, zero valent aqueous sulfur ($S^0_{(aq)}$) is another important redox-active product of sulfate reduction. It can stimulate the formation of multiple thio-As species, acting as a catalyst in redox transformation of As species and can increase As mobility.[23,24] Aqueous sulfide, however, can promote As sequestration into mineral species such as realgar (As_4S_4) and orpiment (As_2S_3), or iron sulfide species like mackinawite.[25,26] Due to these confounding processes, As mobilization in natural aquifers is complicated as a result of sulfur redox reactions. Therefore, the correlation with cryptic sulfur cycling and other geochemical factors associated with arsenic mobilization in groundwater needs further investigation.

In recent years, data-driven methods such as machine learning have been widely applied to understand the geological and hydrogeological factors controlling arsenic mobilization in groundwater in China[27], Italy[28], and India[29]. In Bangladesh, machine learning methods have not been broadly utilized yet to model the spatial distribution of As in groundwater. However, Chakraborty et al. first attempted to develop a probabilistic model of As distribution in the Ganges River delta aquifer system based on anthropogenic activities, geologic, hydrologic, and geomorphic factors through machine learning methods: Random Forest (RF), Logistic Regression (LR), and Boosted Decision Tree (BDT).[17] Tan et al. used a combination of the three methods (RF, LR, and BDT) to improve the accuracy of As risk assessment both in national and regional scales and identify the underlying factors for the spatial distribution of As in groundwater.[30] These studies used different techniques to evaluate the accuracy; however, they did not use the confusion matrix to measure the accuracy of the different classifications of the applied machine learning algorithms. Furthermore, the entire dataset was randomly split for training and testing purposes; therefore, the relationship between the arsenic mobilization and the presence of the commonly identified controlling elements (Fe, Mn, P, SO_4^{2-}) was not well studied yet. Moreover, the application of machine learning to understand the way geochemical factors influence arsenic mobilization in all soil types is still limited. Lastly, the outputs from the tried models were not corroborated with real-world data to enhance confidence and dependability in them.

In response to these research gaps, the present study first attempts to identify the significant controlling features for As mobilization through principal component analysis (PCA), ANOVA F-test feature selection, Chi-square scoring function, and L-index. Second, machine learning algorithms were applied to measure the association between arsenic and its correlated elements along with cryptic sulfur. A confusion matrix has been employed to verify the performance of the methods by measuring their accuracy.

Next, a probabilistic model is developed based on estimated geochemical associations through artificial intelligence techniques to identify the probability of As concentration exceeding to WHO permissible drinking limit. Finally, the outputs are corroborated against spatially interpolated map to identify potential hot zones with reliability. In doing so, the findings of this study are intended to help researchers by producing a statistical model that could be adopted for modeling groundwater arsenic hazards and identifying potential machine learning methods to understand arsenic mobilization in aquifers.

2. Materials and Method

2.1. Study area and data screening

The geochemical dataset from BGS and DPHE was obtained to carry out the analysis and modeling.[31] The dataset consists of detailed information about the geocode, well depth, collection time, and sampling points, which demarcated that groundwater samples collected from all areas irrespective of Bangladesh's geological formation and physiographical properties. The random sampling strategy was applied while collecting groundwater data except for Sundarban and Chittagong hill tracts (Figure 1).

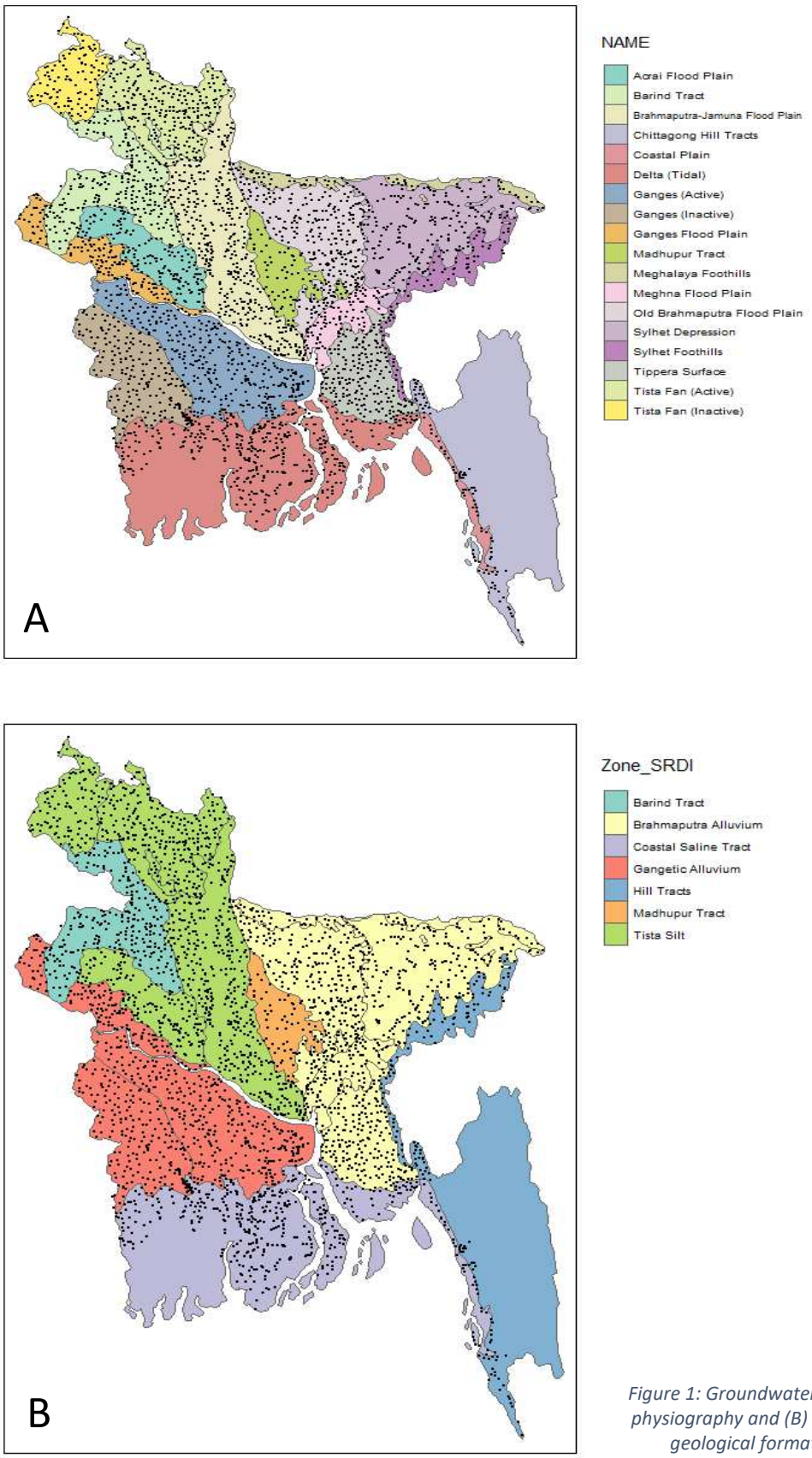


Figure 1: Groundwater sampling points on (A) physiography and (B) soil tracts based on the geological formation of Bangladesh

2.2. Predictor variables

A total of 20 geochemical variables, including arsenic and 19 associated groundwater constituents, were initially considered for statistical association analysis and machine learning model development (Table 1)

Table 1: List of geochemical variables considered in the analysis

Arsenic (As)	Aluminum (Al)	Boron (B)	Barium (Ba)	Calcium (Ca)
Cobalt (Co)	Chromium (Cr)	Copper (Cu)	Iron (Fe)	Potassium (K)
Lithium (Li)	Magnesium (Mg)	Manganese (Mn)	Sodium (Na)	Phosphorous (P)
Silicon (Si)	Sulfate (SO_4^{2-})	Strontium (Sr)	Vanadium (V)	Zinc (Zn)

These geochemical variables were obtained from the BGS and DPHE (BGS/DPHE, 2001) groundwater dataset.[31] The chosen geochemical factors might directly or indirectly influence arsenic mobilization and distribution and were considered for machine learning modeling in a recent study.[30] Therefore, these factors were preliminarily chosen for correlation analysis before arsenic model development. According to their geological formation, the sampling points' geocode was categorized into seven soil tracts (Islam, 1956).

2.3. Principal Component analysis

Several statistical analyses were first applied to identify geochemical variables associated with arsenic occurrence and mobilization. Because the groundwater dataset contained multiple interrelated chemical constituents, PCA was used as an exploratory multivariate technique to reduce dimensionality, minimize redundancy among variables, and identify groups of elements that explain the dominant patterns of groundwater chemistry. In this study, PCA was not used as a standalone predictive model; rather, it served as an initial screening and interpretation tool to support feature identification before subsequent statistical tests and machine learning model development. The PCA procedure included the following preprocessing and analytical steps:

1. Standardization: Since PCA is sensitive to differences in variable scale, all continuous geochemical variables were standardized to zero mean and unit variance before analysis. This step ensured that variables with larger numerical ranges did not disproportionately influence the principal components.

2. Handling Missing Data: Missing values in the continuous variables were treated using mean imputation before conducting PCA. To assess the robustness of this preprocessing step, the PCA results were checked to confirm that the imputation did not substantially alter the resulting component structure or the interpretation of dominant geochemical patterns.

The PCA findings were interpreted in combination with domain knowledge and complementary statistical tests, including correlation analysis, ANOVA F-test, and chi-square scoring. The ANOVA F-test was used to examine whether arsenic occurrence differed significantly with respect to the selected geochemical variables at a significance level of $p < 0.05$. Correlation analysis was performed to evaluate the strength and direction of association between arsenic and other groundwater constituents, while the chi-square test was used to assess whether the observed frequency distributions differed significantly across categorical concentration classes.

PCA was also conducted separately for different geological zones to explore spatial variability in the geochemical controls of arsenic mobilization. The resulting component loadings were examined to identify variables that contributed most strongly to the dominant patterns of groundwater chemistry within each zone. Based on the combined evidence from PCA, correlation analysis, ANOVA F-test, and chi-square scoring, the most influential variables were selected for subsequent machine learning model development using Boosted Decision Tree, Logistic Regression, and Random Forest algorithms.

2.4. Dataset preprocessing

For the machine learning phase, the selected geochemical variables were transformed into binary concentration classes, namely high and low, using threshold values adopted from WHO and EPA guideline values or water-quality criteria. Concentrations equal to or exceeding the corresponding threshold were classified as high, whereas concentrations below the threshold were classified as low. These thresholds were used to define the target and predictor classes for the classification models and are summarized in Table 2.

Table 2: Threshold values used to classify key geochemical variables into low and high concentration classes

Key Component	**Threshold**
Arsenic (As)	10 µg/L
Manganese (Mn)	200 µg/L
Iron (Fe)	300 µg/L
Phosphorus (P)	100 µg/L

Sulfate (SO_4)	250 mg/L

For arsenic (As), a threshold of 10 µg/L was used, following WHO recommendations. Arsenic concentrations above this level were categorized as high impact, while levels below this threshold were classified as low impact. Similarly, for manganese (Mn), concentrations between 200 to 500 µg/L can cause water encrustation, discoloration, and the formation of coatings that may later slough off as a black precipitate. Therefore, in this experiment, Mn concentrations below 200 µg/L were classified as low impact. For iron (Fe), WHO defines concentrations below 300 µg/L as unnoticeable in well water, making this the threshold for categorization. Levels of Fe above this value were classified as high impact. For phosphorus (P), the U.S. Environmental Protection Agency (EPA) limits total phosphorus in flowing waters to 0.1 mg/L. As such, phosphorus levels exceeding 100 µg/L were considered high impact in this study. Lastly, sulfate (SO4) concentrations were evaluated using the WHO recommendation of a taste threshold at approximately 250 mg/L. Levels below this threshold were categorized as low impact, while larger concentrations were classified as high impact.

These thresholds were selected to maintain consistency with internationally recognized drinking-water guidelines and water-quality criteria, while reflecting the potential health, aesthetic, operational, and environmental concerns associated with elevated concentrations of these components. The binary categorization was also intended to improve the interpretability of the machine learning results by defining clear and practically meaningful concentration classes for groundwater quality assessment.

2.5. Machine Learning Model Training

Three supervised machine learning algorithms, Random Forest (RF), Logistic Regression (LR), and Boosted Decision Tree (BDT), were selected to classify groundwater samples into high- and low-impact concentration classes. These algorithms were chosen because they represent different modeling approaches and provide complementary strengths for binary classification. RF is an ensemble tree-based method that combines predictions from multiple decision trees generated from bootstrapped subsets of the data, thereby reducing variance and improving generalization compared with a single decision tree. LR was included as a baseline statistical classification model because it provides an interpretable relationship between predictor variables and the probability of class membership. BDT was used as a boosting-based ensemble method in which sequential decision trees are trained to improve the classification of previously misclassified or poorly predicted samples. The use of these three models allowed comparison between a linear probabilistic classifier and two nonlinear tree-based ensemble methods for predicting arsenic-related groundwater quality classes.

After preprocessing, the selected groundwater components were converted into high- and low-concentration classes based on the threshold values described in Section 2.4. The machine learning models were then trained using three supervised classification algorithms, namely Logistic Regression (LR), Boosted Decision Tree (BDT), and Random Forest (RF), to compare their performance in predicting groundwater quality classes. The analysis focused on two related classification tasks. First, arsenic concentration class was predicted using the selected geochemical components, including Fe, Mn, P, and SO_4, as predictor variables. Second, a reverse classification approach was applied in which arsenic, together with the remaining key components, was used to predict the high- or low-concentration class of each associated component. Geological formation and well-depth information were considered to examine whether the geochemical relationships varied across hydrogeological settings. The dataset was divided into training and testing subsets using an 80:20 split, where 80% of the data were used for model training and the remaining 20% were used for model testing and performance evaluation.

Two complementary classification approaches were adopted in this study. In the first approach, arsenic (As) concentration class was predicted using the key geochemical components iron (Fe), manganese (Mn), phosphorus (P), and sulfate (SO_4) as input features. In the second approach, a reverse classification framework was applied, where As concentration class was used together with the remaining geochemical variables to classify the high- or low-concentration level of each associated component, namely P, Fe, Mn, and SO_4.

For the As classification task, arsenic was used as the target variable, while Fe, Mn, P, and SO_4 were used as predictor variables. Each variable was represented as a binary concentration class, high or low, based on the threshold values described in Section 2.4. The dataset was divided into training and testing subsets using an 80:20 split, and the BDT, LR, and RF models were trained to classify the As concentration class. For the reverse classification task, the same binary class structure was used. In each case, one geochemical component was treated as the target variable, while As and the remaining three components were used as predictor variables. For example, when classifying the P concentration class, As, Fe, Mn, and SO_4 were used as input features. The same modeling procedure and 80:20 train-test split were applied for BDT, LR, and RF to allow consistent comparison across all classification tasks..

2.6. Spatial validation using Kriging interpolation

To spatially evaluate the machine learning outputs, ordinary Kriging interpolation was used to generate continuous arsenic distribution maps from the observed groundwater sampling data. The interpolated maps were then compared with the high-risk zones identified by the machine learning models to examine whether the predicted arsenic hazard patterns were consistent with geostatistical hotspots. This

spatial validation provided an independent basis for assessing the reliability of the model outputs and for identifying areas with elevated groundwater arsenic risk.

Spatial interpolation was used to estimate groundwater arsenic levels in unsampled areas surrounding the observed sampling locations. Ordinary Kriging was performed in ArcGIS to generate continuous spatial surfaces for arsenic and the selected geochemical variables. The interpolated surfaces were then analyzed with reference to SRDI soil tracts to examine the spatial association between arsenic and other groundwater constituents. To quantify these spatial relationships, Lee's L-index was calculated between arsenic and each selected geochemical variable. The L-index measures bivariate spatial association and can be interpreted in a manner similar to Pearson's correlation coefficient, while also incorporating the spatial arrangement of the variables. In addition to the seven SRDI soil tracts, the L-index was computed for the Tippera Surface within the Brahmaputra Alluvium, as classified by the USGS, because this area showed consistently elevated arsenic levels and displayed spatial behavior distinct from the broader Brahmaputra Alluvium.

Ordinary Kriging interpolation was applied to estimate As levels across Bangladesh based on measured concentrations at the sampling locations. Continuous spatial distribution maps for As were generated using the Geostatistical Analyst extension in ArcGIS 10.8. The same interpolation procedure was applied to the selected geochemical variables, including Fe, Mn, and SO_4. The interpolated raster layers were then converted to 1000 m × 1000 m grid cells to provide a common spatial framework for comparison across variables. Lee's L-index, a bivariate spatial association measure, was calculated between As and each selected geochemical variable for individual soil tracts as well as for the whole of Bangladesh. The L-index integrates Pearson's correlation coefficient with Moran's I, allowing both attribute-level association and spatial dependence to be considered in the assessment of As-geochemical relationships.[32,33] Similar to Moran's I, Lee's L-index can be expressed as both local and global statistics. The local L-index was first calculated for each spatial unit to evaluate the bivariate spatial association between As and each selected geochemical variable in relation to neighboring spatial values. The local statistics were then averaged to obtain the corresponding global L-index for the entire study area or for each soil tract. Global L-index values closer to +1 indicate stronger positive spatial association, whereas values closer to −1 indicate stronger negative spatial association. Values near zero suggest weak or no spatial association. Unlike a simple Pearson correlation, Lee's L-index incorporates spatial lag information, allowing the relationship between two variables to be assessed while accounting for their spatial arrangement.[30]

3. Results and Discussion

3.1. Association Analysis

The preliminary association analysis indicated that only a subset of the examined geochemical variables showed notable relationships with arsenic occurrence. Figure 2 presents the correlation coefficients between arsenic concentration and the selected groundwater constituents. Among the analyzed variables, several elements showed positive association with arsenic, although the strength of association varied across components. These results suggest that arsenic occurrence in groundwater is controlled by multiple geochemical interactions rather than by a single dominant factor.

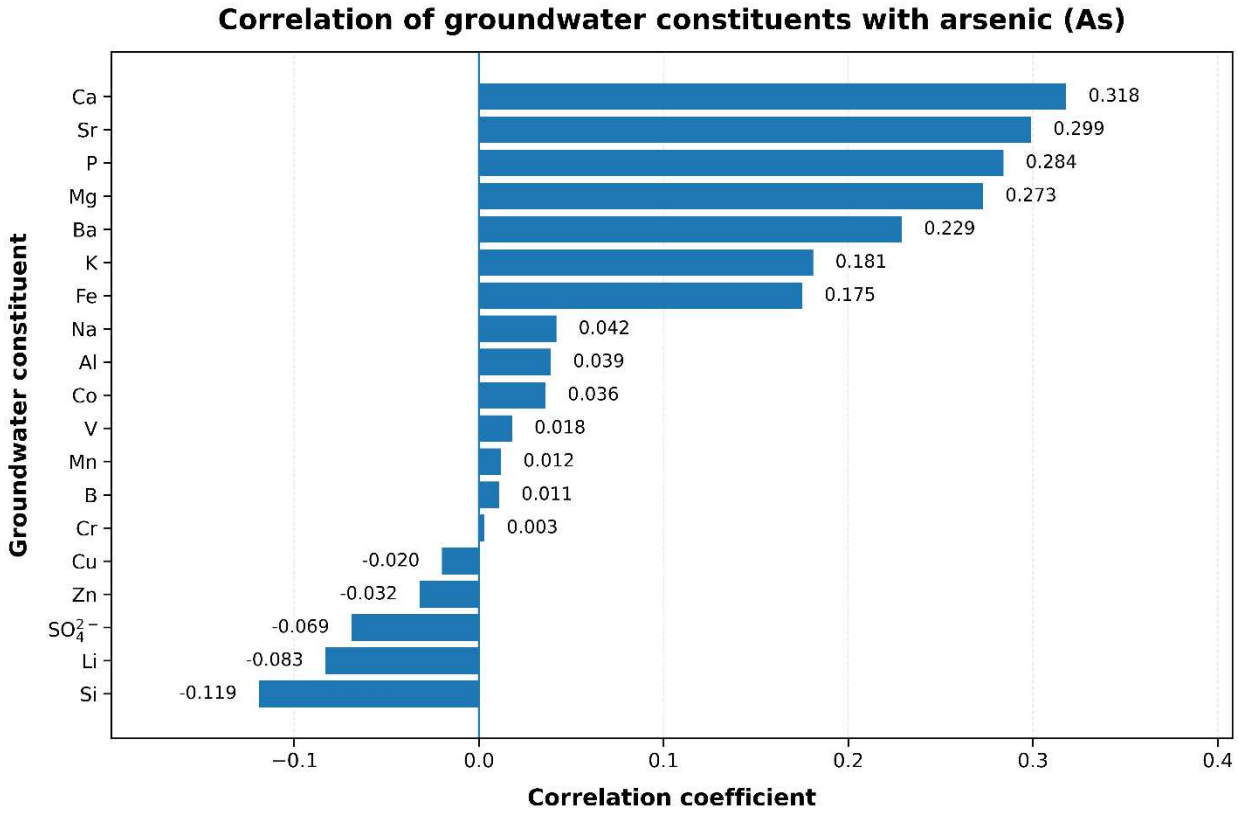


Figure 2: Correlations between Arsenic concentration and other elements

The PCA results further supported the identification of the most influential geochemical variables associated with arsenic occurrence. Based on the PCA and correlation screening, seven variables were identified as having relatively strong associations with arsenic concentration, as shown in Figure 3. The PCA results indicated that the first three principal components captured the dominant patterns of groundwater chemistry. The major contributors to PC1 were Fe, Mn, P, and SO_4, with loadings of 0.82,

0.78, 0.75, and 0.73, respectively. PC2 was mainly associated with Ca, Mg, and Na, while PC3 was influenced by Al, Ba, and V.

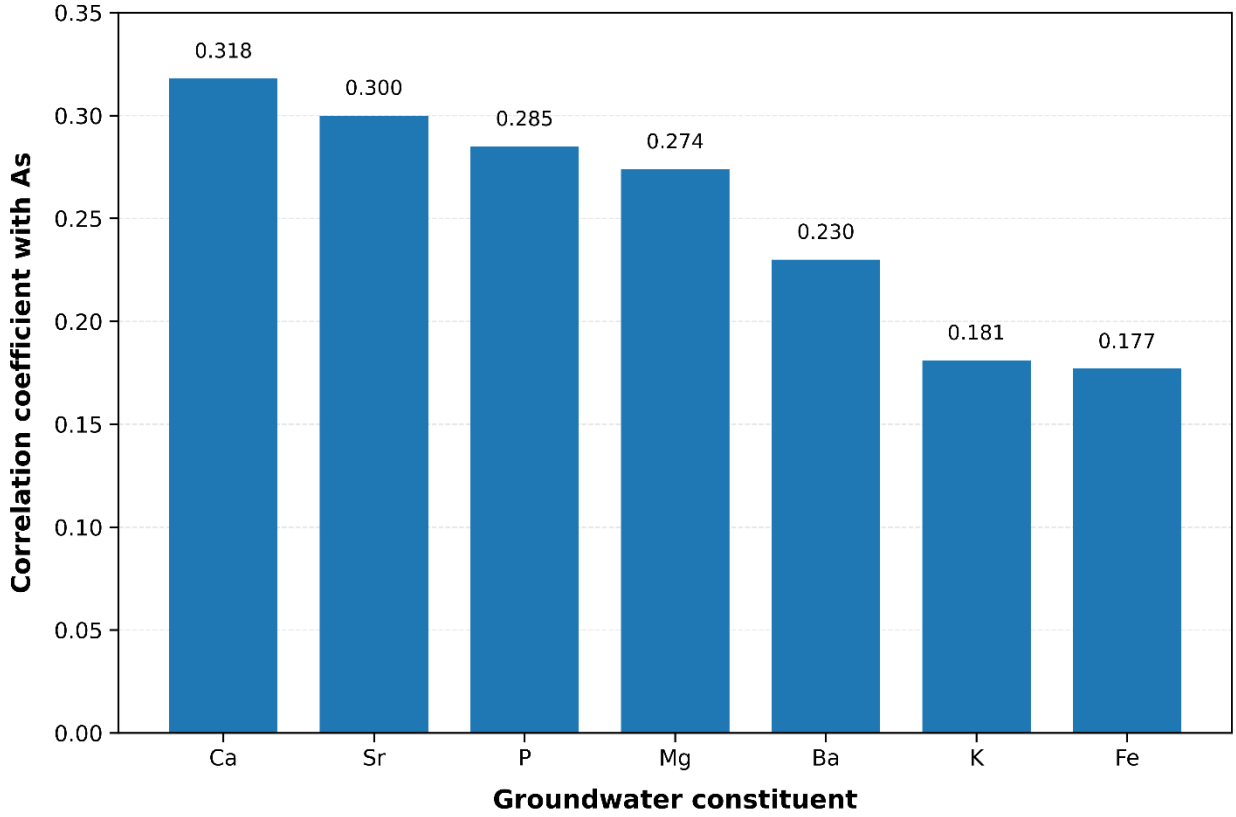


Figure 3: Correlations between Arsenic concentration and the 7 most impactful elements

The PCA results indicated that the first seven principal components collectively explained approximately 85% of the total variance in the geochemical dataset. Among these, the first three principal components accounted for the dominant patterns of groundwater chemistry and were therefore examined in terms of their major variable loadings. Table 3 summarizes the top contributing variables and their corresponding loadings for PC1, PC2, and PC3:

Table 3: The top variables on the first three principal components

Component	Top Contributing Variables	Loadings
PC1	Fe, Mn, P, SO4	0.82, 0.78, 0.75, 0.73
PC2	Ca, Mg, Na	0.76, 0.74, 0.72
PC3	Al, Ba, V	0.71, 0.69, 0.68

PC1 was primarily characterized by high loadings of Fe, Mn, P, and SO_4, indicating that these variables contributed strongly to the dominant geochemical pattern associated with arsenic occurrence. This pattern is consistent with known arsenic mobilization mechanisms involving Fe/Mn redox processes, phosphate competition, and sulfur-related transformations. PC2 was mainly associated with the major cations Ca, Mg, and Na, reflecting broader groundwater chemistry and possible hydrogeochemical evolution. PC3 was influenced by Al, Ba, and V, suggesting that trace or secondary geochemical controls may contribute to localized variation in arsenic occurrence. Overall, the strong contribution of Fe, Mn, P, and SO_4 in PC1 supported their selection as key variables for subsequent machine learning analysis.

The ANOVA F-test was applied to evaluate the statistical association between arsenic occurrence and the selected geochemical variables. Among the seven variables shown in Figure 4, Co, P, and Fe produced the highest F-test values, indicating comparatively stronger discriminatory importance for arsenic classification. Ba and Ca showed moderate F-test scores, whereas Mg and V exhibited relatively lower scores.

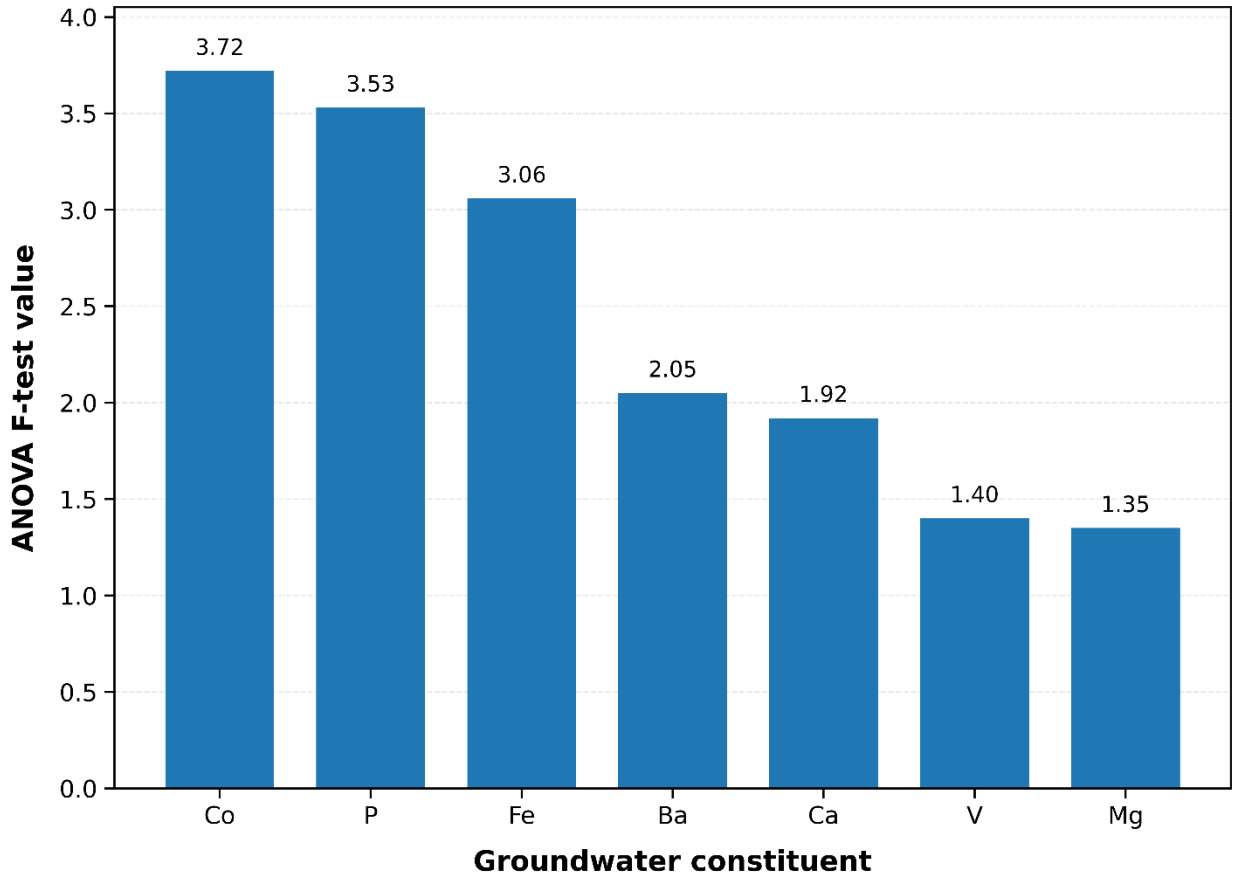


Figure 4: Results of ANOVA F-test showing only 3 out of 7 features have a high impact on Arsenic concentration

The ChiSquare scoring function analysis revealed a strong positive association between abundantly found elements (Na, Ca, K, Al, and Mg) and As concentration. Also, iron, phosphorus, manganese, and sulfate showed a higher positive association with arsenic (Figure 5).

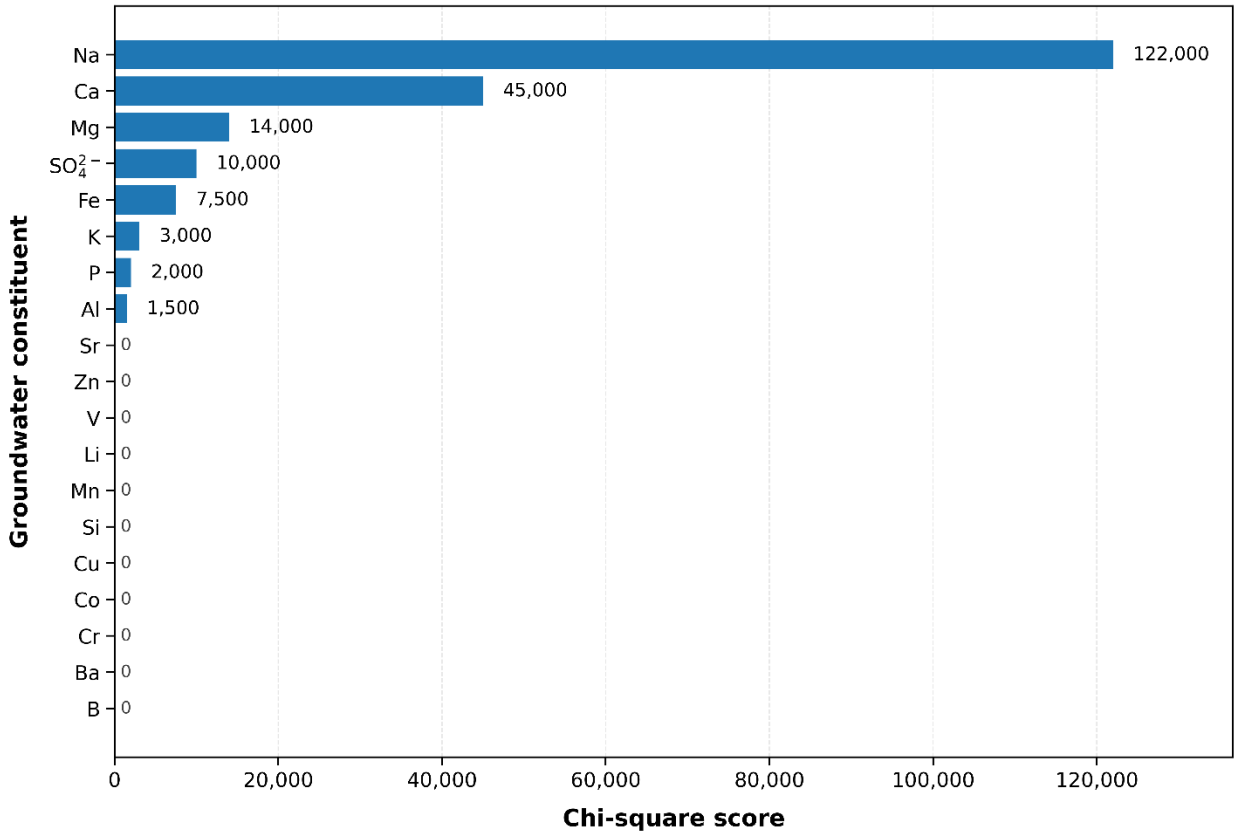


Figure 5: ChiSquare scoring function analysis

3.2. Performance of the machine learning models

The models were trained through finetuning of their hyperparameters. The Parameters used for each of the employed models can be found in Table 4.

Table 4: Finetuning of the hyperparameters for each algorithm

Algorithm	**Parameters**	**Loadings**
BDT	Number of trees, Maximum Number of leaves per tree, Learning rate, and Random seed number	100, 20, 0.2, 15
LR	Random seed number	15

RF	Number of decision trees, Maximum depth of the decision trees, Number of random splits per node	16, 24, 15

After training the machine learning models, we validated the models with the split test data Figure 6 presents the results of predicting arsenic (As) concentration levels using other key geochemical components—iron (Fe), manganese (Mn), phosphorus (P), and sulfate (SO_4)—as input features *(process explained in section 2.5)*.Among all the feature components, iron (Fe) emerged as the most significant predictor component, with a maximum prediction accuracy of 92%. Sulfate, phosphorus, and manganese followed closely, also demonstrating strong predictive capabilities. We can observe that, SO4 had an accuracy of 91%, making it the second most impactful component in As, P and Mn had a significant impact on As with a prediction level of 89%, revealing their importance as major features of the model. Overall, the BDT model outperformed as in most cases, BDT achieved more accurate predictions than the other two algorithms RF and LR, suggesting its superior effectiveness in capturing the complex interactions between arsenic and other groundwater components.

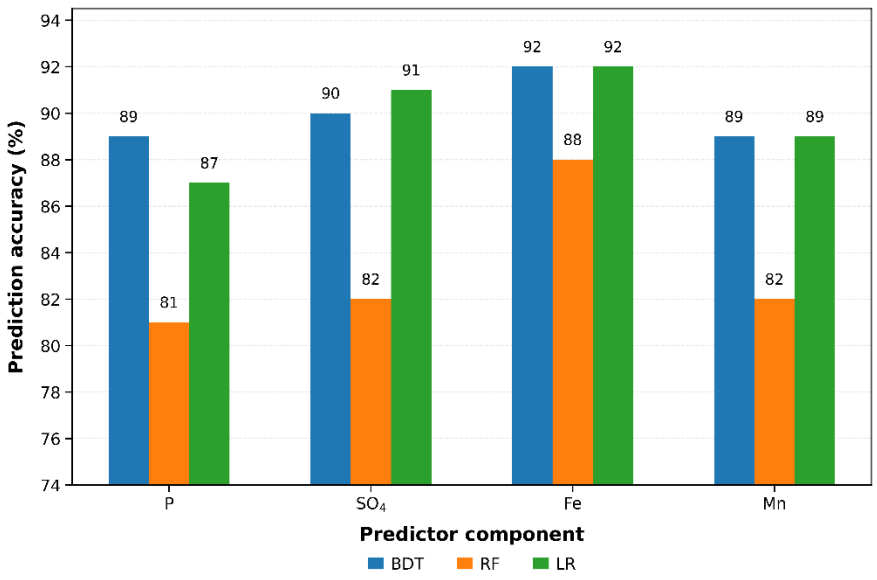


Figure 6: Effects of different components on As level prediction

On the other hand, Figure 7 illustrates the reverse prediction approach, where arsenic (As) concentration was utilized as one of the key predictor features to classify the impact levels of other geochemical components—phosphorus (P), iron (Fe), manganese (Mn), and sulfate (SO_4). The results demonstrated that

the models could accurately classify phosphorus, iron, and manganese levels with accuracies of 88%, 87%, and 85%, respectively, while sulfate was predicted with a slightly lower accuracy of 77%.

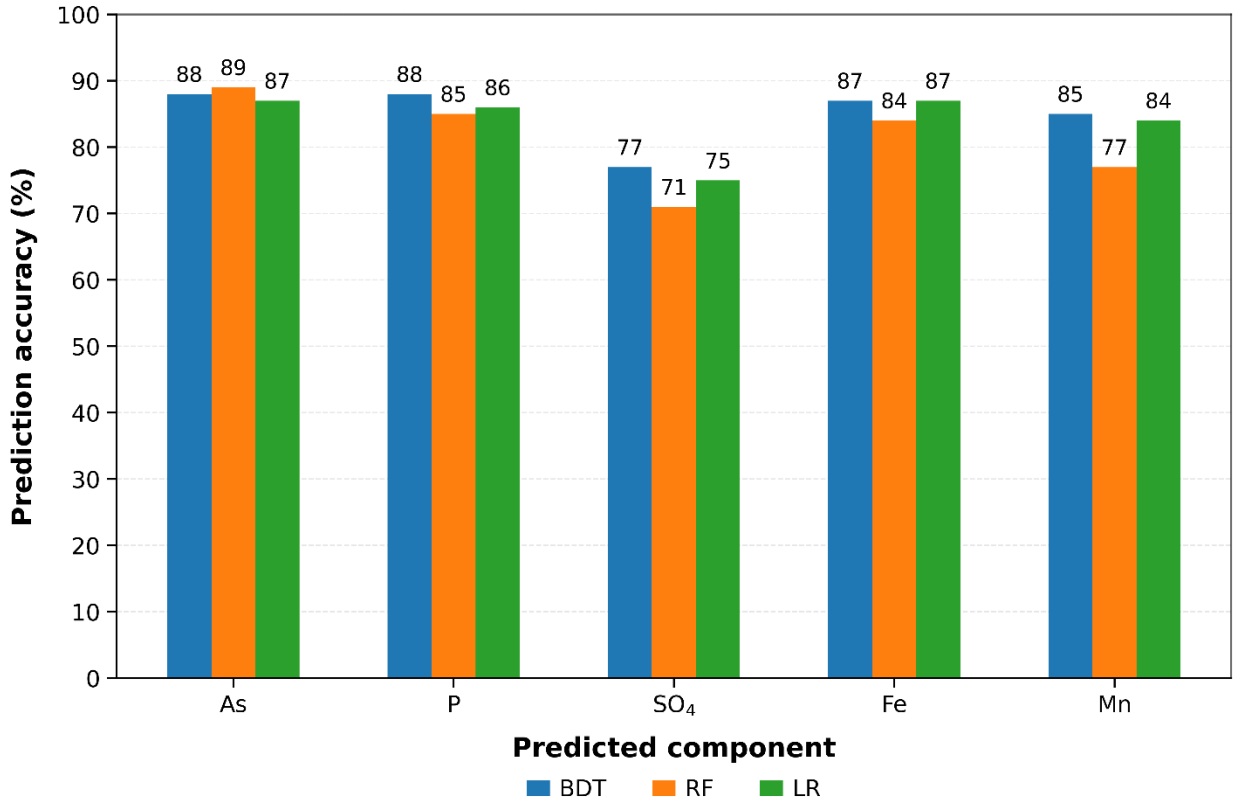
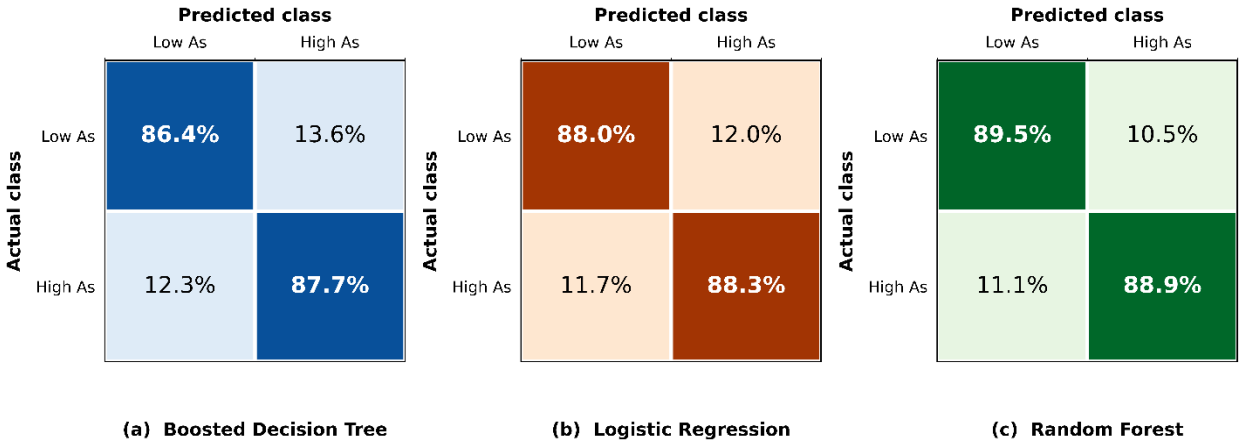


Figure 7: Effects of using As level as a feature on component level prediction

To elaborate, the models were able to predict the level of As with 89% accuracy with different concentrations of P, Mn, Fe, and SO_4 in water. In addition, predicting the concentration of P in water with 88% accuracy can be observed when As, Mn, Fe, and SO_4 are considered as the major important components. These accuracies highlight the strong geochemical associations between arsenic and elements like P, Fe, and Mn, reinforcing their mutual interdependence and significance in groundwater arsenic mobilization. The comparatively lower performance for sulfate indicates a weaker chemical relationship with arsenic.

The overall accuracy of the machine learning model is 85.4%. We can observe the confusion matrices shown in figure 8 to obtain the prediction accuracy of arsenic levels in groundwater for all the three algorithms. In the context of the confusion matrices provided, Class 1 and Class 2 typically represent the

two categories or labels in a binary classification problem. Based on the description of the task, these could represent: Class 1: Groundwater samples with arsenic levels below a certain threshold (e.g., "safe" or "low arsenic levels"); Class 2: Groundwater samples with arsenic levels above the threshold (e.g., "unsafe" or "high arsenic levels").



(a) Boosted Decision Tree **(b) Logistic Regression** **(c) Random Forest**

Figure 8: Confusion metrices of arsenic level prediction accuracy in water considering the impacts of other major component levels.

Below is a detailed table summarizing the performance of the three algorithms—Boosted Decision Tree (BDT), Logistic Regression (LR), and Random Forest (RF)—along with explanations for each metric.

Table 5: Classification result summary including accuracy, precision, recall, and F1-Score for both classes (high and low level of As), along with explanations.

Metric	BDT	LR	RF	Explanation
Accuracy	87.05%	88.15%	89.20%	Accuracy measures the overall percentage of correct predictions (both Class 1 and Class 2). Random Forest has the highest accuracy, indicating better overall performance in case of detecting As level considering the other components in water such as P, Mn, Fe, and SO_4.
Precision (Class 1)	87.54%	88.26%	88.97%	Precision for Class 1 indicates how many of the instances predicted as 'safe arsenic levels' (Class 1) were truly correct. Random Forest has the highest precision, meaning it is slightly better at avoiding false positives for Class 1.
Precision (Class 2)	86.57%	88.04%	89.44%	Precision for Class 2 reflects how many of the instances predicted as 'unsafe arsenic levels' (Class 2) were actually correct. Again, Random Forest outperforms the others, demonstrating better handling of false positives for Class 2.
Recall (Class 1)	86.40%	88.00%	89.50%	Recall for Class 1 represents how many of the actual 'safe arsenic levels' instances (Class 1) were correctly identified by the ML model. Random Forest has the best recall, ensuring fewer false negatives for Class 1.
Recall (Class 2)	87.70%	88.30%	88.90%	Recall for Class 2 indicates how many of the actual 'unsafe arsenic levels' instances (Class 2) were correctly identified. Random

				Forest performs slightly better here, showing its ability to correctly identify more unsafe samples.
F1-Score (Class 1)	86.97%	88.13%	89.23%	F1-Score for Class 1 combines precision and recall, showing a balanced measure of the model's performance. Random Forest performs the best, indicating it balances false positives and false negatives for Class 1 more effectively.
F1-Score (Class 2)	87.13%	88.17%	89.17%	F1-Score for Class 2 is a balanced measure of precision and recall for unsafe arsenic levels. Random Forest performs slightly better, showcasing its reliability in identifying unsafe samples.

We can observe from Figure 8 and Table 5, that Random Forest has the highest Precision and Recall for both Class 1 and Class 2, showing its superior ability to correctly identify both safe and unsafe arsenic levels while minimizing false positives and false negatives. We can see that RF performed more accurately than the other two algorithms predicting the As levels with accuracy of 89%.

3.3. Distribution of iron, manganese, phosphorus, arsenic, and sulfate

The spatial distribution of estimated arsenic (As) concentrations (µg/L), derived using ordinary Kriging interpolation, is presented in Figure 9. The interpolated surface indicates elevated As concentrations predominantly in the central, eastern, and southwestern regions of Bangladesh, encompassing substantial portions of the Brahmaputra Alluvium, Gangetic Alluvium, and Coastal Saline Tract. Localized areas of elevated As concentrations are also observed within the Tista Silt region in northern Bangladesh. Within the Brahmaputra Alluvium, particularly high As concentrations occur in the east-central part of the country, corresponding to the Tippera Surface according to the USGS physiographic classification. Figure 9 further presents the spatial distributions of phosphorus (P), manganese (Mn), iron (Fe), and sulfate (SO_4^{2-}), with sulfate concentrations represented separately for the higher- and lower-concentration ranges. Collectively, these interpolated geochemical distributions provide a spatial basis for examining potential associations between groundwater arsenic occurrence and the major geochemical controls considered in this study.

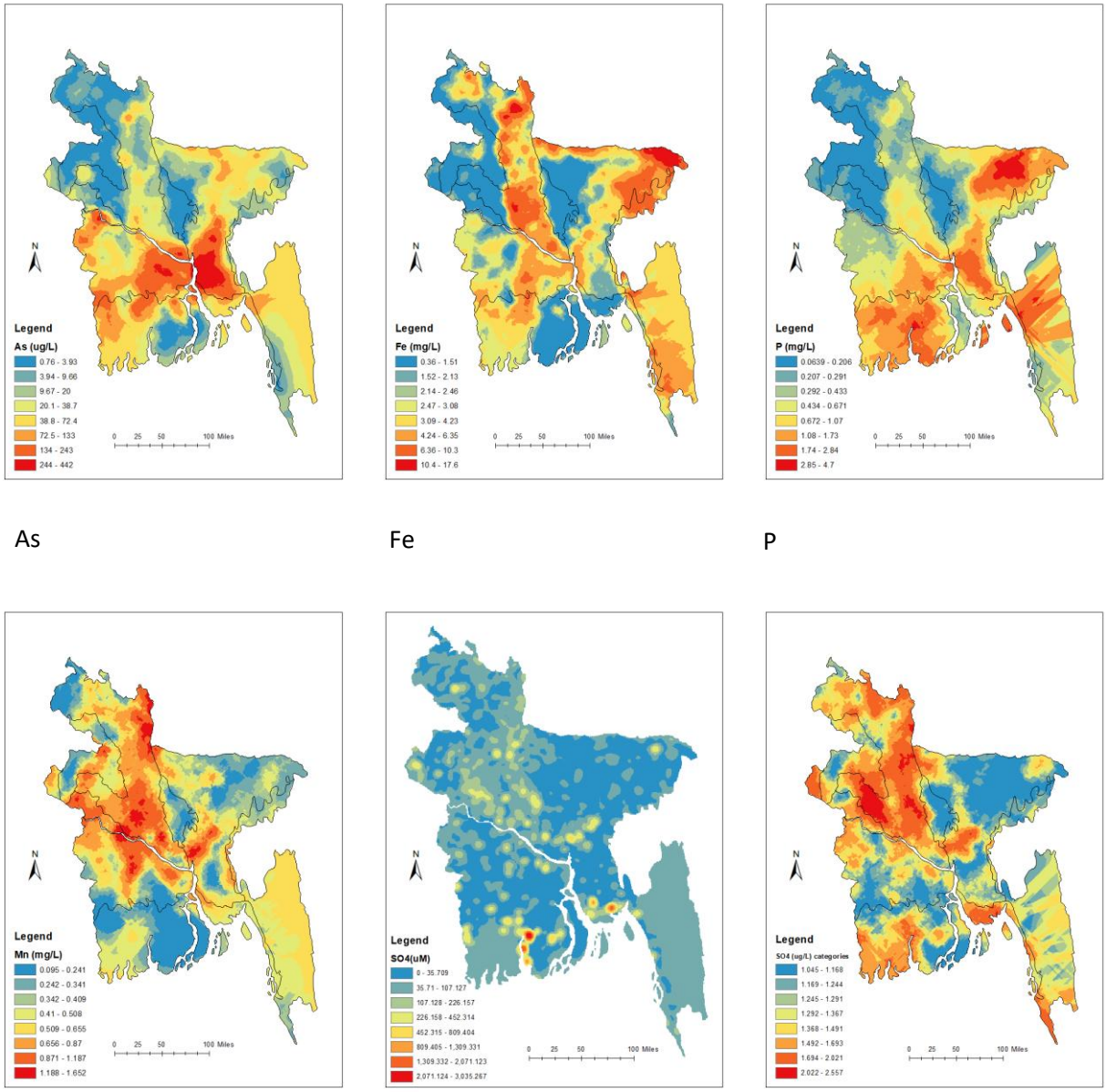
3.5. Validation

A qualitative comparison of the spatial distributions presented in Figure 9 indicates substantial agreement between the Kriging-derived geochemical patterns and the relationships identified by the machine-learning models. Elevated arsenic (As) concentrations generally coincide with areas characterized by high iron (Fe) concentrations. This spatial correspondence is consistent with the model results, in which Fe-based prediction of As levels achieved an accuracy of approximately 92%, suggesting that Fe represents an important predictor of groundwater As occurrence.

A similarly strong spatial correspondence is observed between phosphorus (P) and As. Although the absolute concentration ranges differ, the interpolated spatial patterns of P and As exhibit considerable

similarity across much of the study area. This observation is consistent with the model performance obtained using P as a predictive feature, which yielded an accuracy of approximately 89%. Manganese (Mn) also demonstrates noticeable spatial overlap with As, with elevated Mn occurring in many areas where As concentrations are high. However, several regions exhibit elevated Mn concentrations without corresponding enrichment of As, indicating that the relationship is spatially heterogeneous and that Mn alone may not adequately explain the occurrence of As.

In contrast, the spatial distribution of sulfate (SO_4^{2-}) shows comparatively weaker correspondence with As. Although elevated sulfate concentrations coincide with increased As levels in some locations, this pattern is not consistently observed throughout the study area. Therefore, the spatial results do not provide sufficient evidence to infer a direct or uniform association between elevated sulfate and elevated As concentrations. This comparatively weak spatial agreement is also consistent with the lower predictive performance of sulfate-based models, for which the highest accuracy was approximately 77%. These findings suggest that the relationship between sulfate and As is likely more complex and may depend on interactions with other hydrogeochemical and redox-related processes. Further investigation incorporating multivariate geochemical relationships is therefore required before a definitive association between sulfate concentration and arsenic occurrence can be established.



As Fe P

Mn SO4 High SO4 low

Figure 2: Spatial Kriging Interpolation

410

411 ***3.6. Bivariate relationships***

412 The estimated global bivariate L-index values describing the spatial associations between arsenic (As) and
413 the selected geochemical constituents are presented in Table 6. The results reveal considerable spatial

heterogeneity among the major physiographic regions of Bangladesh, indicating that the relationships between As and individual geochemical constituents vary substantially across different geological and depositional settings.

For iron (Fe), the strongest positive spatial associations with As were observed in the Madhupur Tract (L = 0.806), Gangetic Alluvium (L = 0.744), and Coastal Saline Tract (L = 0.602). Moderate positive associations were also identified in the Tista Silt, Barind Tract, and Tippera Surface. In contrast, the broader Brahmaputra Alluvium exhibited a weak negative association (L = −0.076). Interestingly, the Tippera Surface, located within the Brahmaputra Alluvium in east-central Bangladesh and characterized by comparatively elevated As concentrations, showed a moderate positive As–Fe association (L = 0.418). This contrast between the broader Brahmaputra Alluvium and the Tippera Surface highlights considerable local-scale heterogeneity in the As–Fe relationship. The corresponding local bivariate L-index patterns are illustrated in Figure 10.

Phosphorus (P) exhibited consistently positive spatial associations with As across most physiographic regions. Particularly strong associations were observed in the Tista Silt (L = 0.756), Gangetic Alluvium (L = 0.707), Barind Tract (L = 0.690), Tippera Surface (L = 0.685), and Madhupur Tract (L = 0.655). A moderately positive association was also observed in the Chittagong Hill Tracts (L = 0.591). In comparison, the association was weaker within the broader Brahmaputra Alluvium (L = 0.329), while virtually no spatial association was identified in the Coastal Saline Tract (L = −0.027). At the national scale, P exhibited the strongest positive spatial association with As among the investigated constituents (L = 0.426).

The spatial relationship between manganese (Mn) and As was more variable. The strongest positive association occurred in the Madhupur Tract (L = 0.735), whereas moderate positive associations were observed in the Coastal Saline Tract (L = 0.327) and Chittagong Hill Tracts (L = 0.303). In contrast, negative spatial associations were identified in the Gangetic Alluvium (L = −0.372) and Tippera Surface (L = −0.490). The remaining regions exhibited relatively weak positive associations. At the national scale, the As–Mn association was weakly negative (L = −0.064), suggesting that the relationship between Mn and As is strongly dependent on regional hydrogeochemical conditions rather than being spatially uniform across Bangladesh.

Overall, sulfate (SO_4^{2-}) showed substantially weaker and less consistent spatial associations with As than Fe or P. Most physiographic regions exhibited weak positive or negative L-index values, indicating limited spatial correspondence between sulfate and As concentrations. A notable exception was the Madhupur Tract, where comparatively higher positive associations were observed for both continuous sulfate concentrations (L = 0.375) and the categorized sulfate variable (L = 0.618). At the national scale, however,

the associations remained negative and weak for continuous sulfate (L = −0.062) and somewhat more negative for categorized sulfate (L = −0.231). These results suggest that sulfate-related controls on As occurrence are likely to be region-specific and may depend on interactions with other geochemical and redox processes.

Taken together, the bivariate spatial analysis demonstrates that **Fe and P exhibit the clearest positive spatial associations with groundwater As, although the magnitude and direction of these relationships vary markedly among physiographic regions**. The contrasting regional patterns observed for Mn and sulfate further emphasize that arsenic enrichment in Bangladesh cannot be attributed to a single geochemical factor and is instead governed by spatially heterogeneous hydrogeochemical processes.

Table 6: Global L-index of As and other minerals in different Soil Tracts of Bangladesh (excluding Sundarbans in Coastal Saline Tract and the major part of Chittagong Hill Tract)

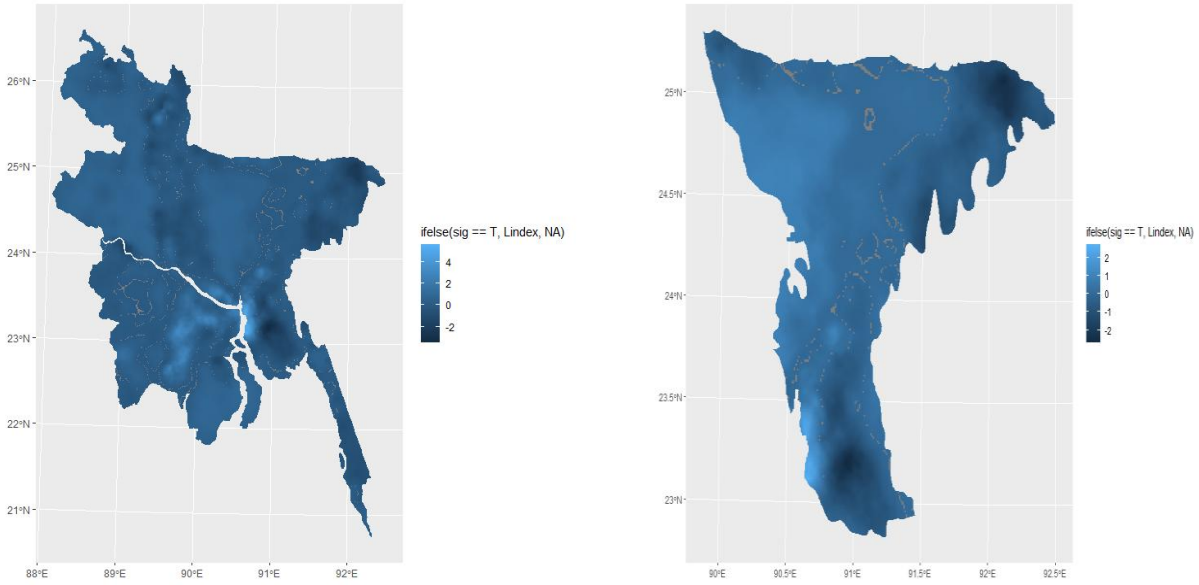


Figure 3: (A) As-Fe local L-index (excluding Sundarbans and Chittagong Hill Tracts) and (B) As-Fe local L-index in Brahmaputra Alluvium

Layer	Fe	P	Mn	SO4	SO4cat
Barind Tract	0.386169	0.69022	0.201274	0.063737	0.014091

Brahmaputra Alluvium	-0.07603	0.328776	0.034963	0.003952	0.03513
Coastal Saline Tract	0.602008	-0.02675	0.326659	-0.09598	-0.05074
Gangetic Alluvium	0.744027	0.706503	-0.37199	-0.12347	-0.31969
Chittagong Hill Tracts	-0.16785	0.590971	0.30271	0.269525	-0.12443
Madhupur Tract	0.80609	0.65491	0.735387	0.374986	0.617645
Tista Silt	0.416751	0.755682	0.155443	-0.06176	-0.19629
Tippera Surface (within Brahmaputra Alluvium)	0.418186	0.684766	-0.49023	-0.1697	-0.25481
Whole Bangladesh	0.157172	0.425823	-0.06429	-0.06238	-0.23059

4. Conclusion

This study developed an integrated statistical, machine-learning, and geospatial framework to assess the geochemical factors associated with groundwater arsenic (As) hazard in Bangladesh. By combining principal component analysis (PCA), ANOVA F-test and Chi-square-based feature assessment, supervised machine-learning models, ordinary Kriging interpolation, and bivariate spatial analysis using Lee's L-index, the study addressed an important limitation of earlier approaches that considered arsenic prediction and geochemical associations largely in isolation. The results demonstrate that groundwater As occurrence is associated with a combination of geochemical factors and that the strength and direction of these relationships vary considerably across the major physiographic regions of Bangladesh.

The statistical analyses identified Fe, Mn, P, and SO_4 among the principal geochemical variables associated with arsenic occurrence. In particular, Fe, Mn, P, and SO_4 contributed strongly to the dominant geochemical pattern represented by the first principal component, while the ANOVA analysis further highlighted Fe and P among the variables with comparatively high discriminatory importance for As classification. The machine-learning results confirmed that these geochemical relationships contain substantial predictive information. Among the individual predictors, Fe provided the highest maximum accuracy for As-level classification at approximately 92%, while P and Mn achieved accuracies of approximately 89%; sulfate also demonstrated predictive capability, although its spatial relationship with As was less consistent. When Fe, Mn, P, and SO_4 were considered collectively for As classification, Random Forest produced the best overall performance, with an accuracy of 89.20% and consistently strong precision, recall, and F1-scores for both low- and high-As classes. These results demonstrate the utility of machine-learning approaches for distinguishing groundwater samples with elevated As concentrations using readily available geochemical information.

The geospatial analyses provided an important additional level of evaluation. Ordinary Kriging identified elevated As concentrations predominantly across the central, eastern, and southwestern parts of Bangladesh, including substantial portions of the Brahmaputra Alluvium, Gangetic Alluvium, and Coastal Saline Tract, together with localized enrichment in the Tista Silt. Particularly elevated concentrations were also observed in the Tippera Surface of east-central Bangladesh. Comparison of the interpolated surfaces showed substantial spatial correspondence of As with Fe and P, supporting the relationships identified by the statistical and machine-learning analyses. In contrast, Mn exhibited greater spatial heterogeneity, whereas sulfate showed a comparatively weak and inconsistent spatial correspondence with As, indicating that sulfur-related controls are likely more complex and dependent on interactions with other hydrogeochemical and redox processes.

The bivariate spatial analysis further demonstrated that these geochemical relationships cannot be regarded as spatially uniform. For example, strong positive As–Fe associations were observed in the Madhupur Tract, Gangetic Alluvium, and Coastal Saline Tract, whereas the broader Brahmaputra Alluvium showed a weak negative association. Within the latter, however, the Tippera Surface exhibited a moderately positive As–Fe relationship. Phosphorus showed the strongest positive association with As at the national scale and consistently positive associations across most physiographic regions, whereas Mn and sulfate displayed considerably greater regional variability. These findings emphasize that arsenic hazard in Bangladesh is governed by spatially heterogeneous hydrogeochemical conditions rather than by a single geochemical control.

Overall, the study addresses the identified research gaps by integrating geochemical feature identification, quantitative machine-learning assessment, confusion-matrix-based model evaluation, spatial interpolation, and region-specific bivariate analysis within a common framework. The agreement between the statistical, machine-learning, and geospatial results strengthens confidence in the identification of Fe and P as particularly important indicators of groundwater As occurrence, while also demonstrating that Mn and sulfate relationships are more context dependent. The proposed framework may therefore provide a useful basis for identifying arsenic-prone areas, prioritizing groundwater monitoring, and supporting risk-informed groundwater management in Bangladesh. Nevertheless, the identified relationships should be interpreted primarily as statistical and spatial associations rather than direct evidence of causality. Future studies incorporating temporally resolved groundwater observations, aquifer depth and sediment characteristics, redox conditions, organic matter, arsenic speciation, and additional hydrogeological variables could further clarify the mechanisms governing arsenic mobilization and improve the transferability and predictive robustness of the proposed framework.

References

1. P.L. Smedley DGK. A review of the source, behaviour and distribution of arsenic in natural waters. *Appl Geochemistry*. 2002;17(5):517–568. doi:https://doi.org/10.1016/S0883-2927(02)00018-5

2. Flanagan S V., Johnston RB, Zheng Y. Arsenic in tube well water in Bangladesh: health and economic impacts and implications for arsenic mitigation. *Bull World Health Organ*. 2012;90(11):839-846. doi:10.2471/BLT.11.101253

3. Smith A, Lingas E, Rahman M. Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. *Bull World Health Organ*. 2000;78(9):1093-1103. doi:doi:10.1590/S0042-96862000000900005

4. Hassan KM, Fukushi K, Turikuzzaman K, Moniruzzaman SM. Effects of using arsenic-iron sludge wastes in brick making. *Waste Manag*. 2014;34(6):1072-1078. doi:10.1016/j.wasman.2013.09.022

5. Rahman MA, Jalil MA, Ali MA. Transformation of arsenic in the presence of cow dung and arsenic sludge disposal and management strategy in Bangladesh. *J Hydrol*. 2014;518(PC):486-492. doi:10.1016/j.jhydrol.2013.05.005

6. Safiuddin M, Karim MM. Groundwater Arsenic Contamination in Bangladesh: Causes, Effects and Remediation. *Proc 1st IEB Int Conf 7th Annu Pap Meet*. 2001;(September):14. doi:10.1080/10934520701566744

7. Reza AHMS, Jean JS, Lee MK, et al. Implications of organic matter on arsenic mobilization into groundwater: Evidence from northwestern (Chapai-Nawabganj), central (Manikganj) and southeastern (Chandpur) Bangladesh. *Water Res*. 2010;44(19):5556-5574. doi:10.1016/j.watres.2010.09.004

8. Tondel M, Rahman M, Magnuson A, Chowdhury IA, Faruquee MH, Ahmad SA. The relationship of arsenic levels in drinking water and the prevalence rate of skin lesions in Bangladesh. *Environ Health Perspect*. 1999;107(9):727-729. doi:10.1289/ehp.99107727

9. Nriagu J, Bhattacharya P, Mukherjee A, Bundschuh J, Zevenhoven R, Loeppert R. Arsenic in soil and groundwater: an overview. *Arsen Soil Groundw Environ - Biogeochem Interact Heal Eff Remediat*. 2007;9(06):3-60. doi:10.1016/s0927-5215(06)09001-1

10. Mukherjee J, Wisniwski S. Bangladesh - Arsenic Mitigation - Water Supply Project. Published online 1998:566. http://agris.fao.org/agris-search/search.do?recordID=US2012418648

11. UNICEF. Policy brief June 2018: Mitigating Arsenic in drinking water. 2018;(June):1-6.

12. Hug SJ, Winkel LHE, Voegelin A, Berg M, Johnson AC. Arsenic and Other Geogenic Contaminants in Groundwater-A Global Challenge. *Chimia (Aarau)*. 2020;74(7):524-537. doi:10.2533/CHIMIA.2020.524

13. Wang S, Mulligan CN. Effect of natural organic matter on arsenic release from soils and sediments into groundwater. *Environ Geochem Health*. 2006;28(3):197-214. doi:10.1007/s10653-005-9032-y

14. Bhattacharya P, Hasan MA, Sracek O, et al. Groundwater chemistry and arsenic mobilization in the Holocene flood plains in south-central Bangladesh. *Environ Geochem Health*. 2009;31(SUPPL. 1):23-43. doi:10.1007/s10653-008-9230-5

15. Bhattacharya P, Jacks G, Frisbie S, Euan S, Naidu R, Sarkar B. *Arsenic in the Environment: A Global Perspective*.; 2002.

16. Anawar HM, Akai J, Komaki K, et al. Geochemical occurrence of arsenic in groundwater of Bangladesh: Sources and mobilization processes. *J Geochemical Explor*. 2003;77(2-3):109-131. doi:10.1016/S0375-6742(02)00273-X

17. Chakraborty M, Sarkar S, Mukherjee A, et al. Modeling regional-scale groundwater arsenic hazard in the transboundary Ganges River Delta, India and Bangladesh: Infusing physically-based model with machine learning. *Sci Total Environ*. 2020;748:141107. doi:10.1016/j.scitotenv.2020.141107

18. Gnanaprakasam ET, Lloyd JR, Boothman C, et al. Microbial community structure and arsenic biogeochemistry in two arsenic-impacted aquifers in Bangladesh. *MBio*. 2017;8(6). doi:10.1128/mBio.01326-17

19. Nickson RT, Mcarthur JM, Ravenscroft P, Burgess WG, Ahmed KM. Mechanism of arsenic release to groundwater, Bangladesh and West Bengal. *Appl Geochemistry*. 2000;15(4):403-413. doi:10.1016/S0883-2927(99)00086-4

20. Smith E, Naidu R, Alston AM. Chemistry of Inorganic Arsenic in Soils: II. Effect of Phosphorus, Sodium, and Calcium on Arsenic Sorption. *J Environ Qual*. 2002;31(2):557-563. doi:10.2134/jeq2002.5570

21. Canfield DE, Stewart FJ, Thamdrup B, et al. A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. *Science (80-)*. 2010;330(6009):1375-1378. doi:10.1126/science.1196889

22. Guo H, Zhou Y, Jia Y, et al. Sulfur cycling-related biogeochemical processes of arsenicmobilization in the western hetao basin, China: Evidence from multiple isotope approaches. *Environ Sci Technol*. 2016;50(23):12650-12659. doi:10.1021/acs.est.6b03460

23. Couture RM, Van Cappellen P. Reassessing the role of sulfur geochemistry on arsenic speciation in reducing environments. *J Hazard Mater*. 2011;189(3):647-652. doi:10.1016/j.jhazmat.2011.02.029

24. Burton ED, Johnston SG, Planer-Friedrich B. Coupling of arsenic mobility to sulfur transformations during microbial sulfate reduction in the presence and absence of humic acid. *Chem Geol*. 2013;343:12-24. doi:10.1016/j.chemgeo.2013.02.005

25. Aziz Z, Bostick BC, Zheng Y, et al. Evidence of decoupling between arsenic and phosphate in shallow groundwater of Bangladesh and potential implications. *Appl Geochemistry*. 2017;77:167-177. doi:10.1016/j.apgeochem.2016.03.001

26. Buschmann J, Berg M. Impact of sulfate reduction on the scale of arsenic contamination in groundwater of the Mekong, Bengal and Red River deltas. *Appl Geochemistry*. 2009;24(7):1278-1286. doi:10.1016/j.apgeochem.2009.04.002

27. Wu J, Teng Y, Chen H, Li J. Machine-learning models for on-site estimation of background concentrations of arsenic in soils using soil formation factors. *J Soils Sediments*. 2016;16(6):1787-1797. doi:10.1007/s11368-016-1374-9

28. Dalla Libera N, Pedretti D, Tateo F, Mason L, Piccinini L, Fabbri P. Conceptual Model of Arsenic Mobility in the Shallow Alluvial Aquifers Near Venice (Italy) Elucidated Through Machine Learning and Geochemical Modeling. *Water Resour Res*. 2020;56(9):1-20. doi:10.1029/2019WR026234

29. Podgorski J, Wu R, Chakravorty B, Polya DA. Groundwater arsenic distribution in india by machine learning geospatial modeling. *Int J Environ Res Public Health*. 2020;17(19):1-17. doi:10.3390/ijerph17197119

30. Tan Z, Yang Q, Zheng Y. Machine Learning Models of Groundwater Arsenic Spatial Distribution in Bangladesh: Influence of Holocene Sediment Depositional History. *Environ Sci Technol*. 2020;54(15):9454-9463. doi:10.1021/acs.est.0c03617

31. UNICEF. Bangladesh national drinking water quality survey of 2009. Dhaka, Bangladesh: Bangladesh Bureau of Statistics. *Bangladesh Bur Stat Plan Div Minist Plan Gov People's Repub Bangladesh*. Published online 2009:105. https://washdata.org/sites/default/files/documents/reports/2019-06/Bangladesh-2009-MICS-water-quality-report.pdf

32. Lee S. Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. *J Geogr Syst*. Published online 2001:369-385.

33. Moran PAP. The Interpretation of Statistical Maps. *J R Stat Soc*. 1948;10(2):243-251. http://www.jstor.org/stable/2983777

Author Contributions:

Dr. Tasnia Ashrafi Heya has contributed in data curation and analyses with PCA and machine learning. Dr. Md Shakil Kashem led geospatial and L-index analyses. Tahsin Habib contributed in drafting the manuscript. Dr. Shubhra Bhattacharjee and Dr. Asef Mohammad Redwan combinedly conceptualized the manuscript. Dr. Tanvir Mustafy contributed in editing and reviewing. Dr. Asef Mohammad Redwan supervised the work.

Funding:

No funding was received for this manuscript.

Competing Interests:

The author(s) declare no competing interests.

Data Availability Statement:

For this study, authors used the following two Bangladesh Government databases:

1. Department of Public Health Engineering (DPHE): https://dphe.gov.bd/pages/static-pages/6922df45933eb65569e20fbd

2. Bangladesh Water Development Board (BWDB) Database: https://hydrology.bwdb.gov.bd/index.php?pagetitle=ground_water_daily_data_available&sub4=226&SUBID=131&id=125&id2=127&id3=309