ORIENTAL JOURNAL OF CHEMISTRY
An International Open Access, Peer Reviewed Research Journal

www.orientjchem.org

ISSN: 0970-020 X
CODEN: OJCHEG
2026, Vol. 42, No.(1):
Pg. 194-201

Comparative Analysis of Random Forest and Decision Tree for Arsenic Prediction in Groundwater in Begusarai, Bihar

ABHISHEK KUMAR MISHRA[1], SURAJ KUMAR[2], ASTHA KUMARI[3], MADAN CHANDRA MAURYA[4] NITYANAND SINGH MAURYA[5]*

[1,2,5]Department of Civil Engineering, National Institute of Technology Patna, Bihar–800005, India.
[3]Department of Civil Engineering, Bakhtiyarpur College of Engineering, Bihar-803212, India.
[4]Civil Engineering Department, Madan Mohan Malaviya University of Technology, Gorakhpur, Uttar Pradesh-273010, India.
*Corresponding author E-mail-nsmaurya@nitp.ac.in

http://dx.doi.org/10.13005/ojc/420121

(Received: May 30, 2025; Accepted: August 07, 2025)

ABSTRACT

A comprehensive statistical analysis was conducted on groundwater samples from Matihani and Cheria Bariarpur blocks of Begusarai District, Bihar to assess physicochemical properties and heavy metal concentrations. While pH levels in both blocks remained within acceptable limits (6.5–8.5), arsenic (As) concentrations significantly exceeded the permissible limit (0.01 mg/L), with mean values of 0.129 mg/L in Matihani and 0.1656 mg/L in Cheria Bariarpur. Machine learning models-Decision Tree Regressor (DTR) and Random Forest Regressor (RFR)-were applied to predict arsenic concentrations. RFR outperformed DTR, achieving a lower MSE (0.001867), RMSE (0.043208), and higher NSE and R^2 values (0.895), demonstrating superior predictive accuracy. The results indicate that arsenic contamination poses a serious public health risk, and Random Forest offers a robust tool for spatial prediction and water quality management. This study highlights the importance of integrating AI-driven modeling with field data to enhance groundwater risk assessment and support informed mitigation strategies.

Keywords: Arsenic, Decision tree, Machine learning, Modelling, Prediction.

INTRODUCTION

For the Earth to remain green and habitable, water is necessary. Effective management of water resources is essential to every country's progress. The UN General Assembly acknowledged access to clean drinking and recreational water in 2008[1]. Consuming water tainted with bacteria, viruses, and heavy metals may spread a number of illnesses with high death and morbidity rates, including cholera, dysentery, diarrhea, skin cancer, and typhoid. Since these illnesses make up 80% of all diseases that have been discovered so far, they have important[2]. In Bihar's cities and rural areas, groundwater is a vital resource for residential, commercial, and agricultural uses[3]. Due to considerable industrial expansion and population increase, the need for water is increasing daily[4,5]. However, one of the biggest environmental problems facing any country today

This is an Open Access article licensed under a Creative Commons license: Attribution 4.0 International (CC- BY).
Published by Oriental Scientific Publishing Company © 2018

is the supply of fresh water, which is a result of several human activities. One of the main concerns is groundwater contamination[6]. Groundwater quality is significantly impacted, particularly in India, by the discharge of untreated sewage, poor solid waste management, widespread use of pesticides and fertilizers in agriculture, overexploitation of groundwater, and changes in land use and land cover[7,8]. The geography of the area, soil characteristics, groundwater and rock interactions, and climate may all have an impact on groundwater quality in addition to pollution[9]. In light of the previously described elements, it is now essential to assess the quality of groundwater and carry out corrective. Contamination of drinking water may happen during treatment and transportation processes or at the source as a result of heavy metals and minerals (geogenic). With careful monitoring and targeted water quality management techniques, the latter element may be controlled. However, when naturally occurring heavy metal concentrations above allowable limits, pollution becomes a significant problem. About seventeen of the fifty heavy metals are known to cause cancer, with lead, arsenic, mercury, cadmium, and vanadium being the most hazardous[10]. Anthropogenic and natural sources of heavy metals seep into groundwater, causing environmental issues around the globe. One dangerous water contaminant is arsenic (As)[11]. Because arsenic is a trace element and may be released by geogenic or human activity, it is important to identify and quantify the hazards of consuming it through the skin, inhalation, and ingestion. Arsenic exhibits four oxidation states, which are designated as arsine, arsenic, arsenite, and arsenate. As(III) and As(V) are more commonly found in groundwater and surface water, respectively. Its solubility depends on the pH and ionic environment[12]. In general, As(V) is less hazardous than As(III), and organic forms of arsenic are less dangerous than inorganic ones[13]. Arsenic oxides are recognized to have the potential to be more hazardous than other arsenic compounds. Even when taken at smaller amounts over longer periods of time, many substances have a higher toxicity index[14]. Significant non-carcinogenic hazards, including gangrene, keratosis, hyperpigmentation, black foot disease, and vascular disorders, as well as carcinogenic risks, such cancer, have been linked to long-term chronic exposure to inorganic arsenic[15]. The International Agency for Research on Cancer (IARC) has designated arsenic as a class I human carcinogenic metalloid due to the potentially fatal effects of exposure[16]. As per IS 10500:2012 permissible limit of arsenic is 0.01 mg/L[17]. Because of this, arsenic is currently regarded as a dangerous material worldwide, posing a number of immediate and long-term health risks. The incidence of diseases like cancer has increased significantly. A number of things may infect humans with arsenic, including food, beverages, and inadvertently[18]. Drinking water is considered a significant exposure route for arsenic when compared to other ingestible channels, such as cutaneous and inhalation. Since arsenic in India's arsenic-rich groundwater is causing increasing worries about both human health and the environment, it is imperative to evaluate the associated health risks[19].

Rapid urbanization and excessive water usage are now occurring in the research locations that were chosen. Understanding the relationships between arsenic levels and other medium characteristics is thus crucial. Since it is highly difficult to acquire and quantify arsenic samples, a variety of machine learning models were created, and their effectiveness in predicting arsenic concentrations based on other input characteristics was evaluated.

MATERIAL AND METHODS

Study Area

One of India's most fertile and highly inhabited areas is the Gangetic Plains. There are over 83 million people residing in the Middle and partially Upper Ganga Plains of Bihar, which has an area of about 94,163 km^2. The Mid-Gangetic Plains and the floodplains of the Sone and Gandak rivers, two of its tributaries, were the study's locations. To achieve this, samples were drawn at random from tube wells close to and distant from the Ganga River's banks at locations on both banks. Samples were specifically taken from Begusarai, with an emphasis on the blocks of Matihani and Cheria Bariarpur. Fig. 1 below shows the map that describes the research area:

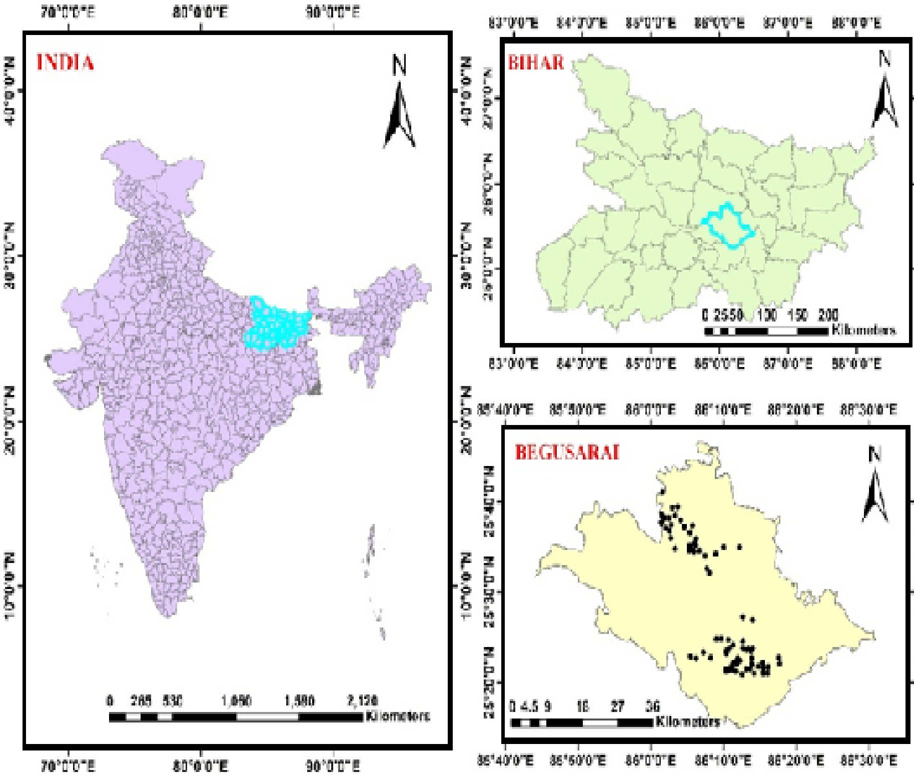

Fig. 1. Study area of Begusarai (Cheria Bariarpur and Matihani)

Geology and Hydrogeology

The district of Begusarai is located on the Ganga River's northern bank. This district is traversed by the Burhi Gandak, Balan, Bainty, Baya, and Chandrabhaga rivers. The district is 1,918 square kilometers in size and is home to 2,954,367 people. Rainfall in the area averages 138.4 cm per year. Temperatures may dip to as low as 8°C in the winter, while they can rise as high as 40°C in the summer. The primary source for the people living in the Begusarai area for agricultural and domestic purposes is groundwater. For this research, the blocks of Cheria Bariarpur and Matihani were chosen in order to measure the amounts of arsenic in regions both close to and far from the Ganga River's banks. While Cheria Bariarpur block is close to the rivers Burhi Gandak, Balan, Bainty, Baya, and Chandrabhaga, Matihani block is situated on the Ganga River's bank.

Sample collection and analysis

From the tube wells (depth 60-300 feet) in the Begusarai district (Matihani and Cheria Bariarpur Block), a total of 100 samples were taken. (Fig. 1). The sample protocol outlined in[20] was adhered to. Prior to sampling, the tube well's stagnant water was eliminated. Groundwater samples were collected using pre-cleaned polyethylene bottles that had been rinsed with fresh water and demineralized water after being cleaned with a 10% nitric acid solution. Prior to sample collection, the sampling vials were washed with the sample water. EUTECH's multi-parameter PCSTestr 35 was used on-site to measure the sample's pH. To determine the presence of heavy metals, samples were taken from each tube well and acidified using concentrated HNO3 to bring the pH down to 2. After being collected, the samples were placed in an ice box and brought to the lab, where they were maintained at 4°C. Prior to laboratory examination, samples were filtered via a 0.45-μm pore-size membrane. An Agilent 5110 ICP-OES inductively coupled plasma emission spectrometer was used to measure the levels of heavy metals (iron and arsenic).

Model Development

Scaling or Normalization

When the input data has different scales, machine learning algorithms often perform badly[21].

Consequently, Eqn. 1 was used to scale the input variables (from 0 to 1):

$$X_{Scale} = \frac{X - X_{min}}{X_{max} \quad X_{min}} \tag{1}$$

Thirty percent of the dataset was put aside for testing, while seventy percent was used to train the machine learning models in order to adequately assess the models. The arsenic prediction was tested against the observed values after measured water quality parameters, aside from the quantity of arsenic, were specified as input variables. Python was used to implement the different models. The theoretical ideas and outcomes of the varoius machine learning models are shown and discussed below.

Random forest regressor model (RFR)

Breiman (2001) created the Random Forest Regressor Model (RFR), an ensemble technique that builds many decision trees using random sampling with replacement for making repetitive target variable prediction. The method is presented in Fig. 2 and is explained by[21], utilizes the performance of many decision tree algorithms to predict the level of arsenic. A training subset chosen at random with replacement is used by each decision tree, and this subset is repeated as many times as there are trees in the ensemble. A final forecast is then generated by combining the results of various decision trees[22]. Using bootstrap sampling, each tree is constructed using a random subsample of the training dataset; the samples that are excluded from this subsample are known as "Out of Bag" (OOB) samples[23]. The RFR model is internally cross-validated using these OOB samples[24]. Hyperparameter tweaking was used to improve the prediction of arsenic concentration levels after the first RFR model was developed.

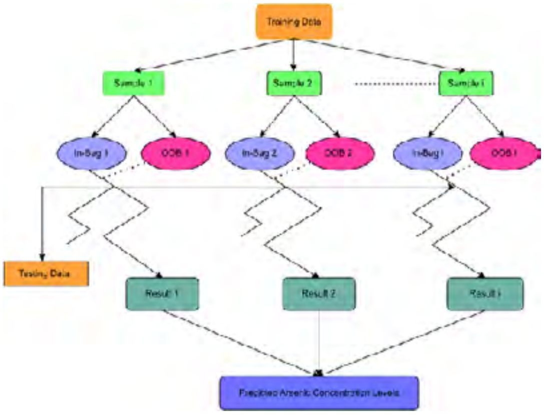

Fig. 2. Flow chart of the Random Forest regressor algorithm

Decision Tree regressor model (DTR)

Regression and classification issues are addressed using non-parametric machine learning methods known as decision trees (DTs). Unlike black-box algorithms, DTs feature an easy-to-understand decision-making process and are very intuitive[21]. Before rendering decisions based on a range of input variables organized into layers of decision branches, the algorithm begins at a root node and proceeds via internal and terminal nodes. Following the first split of the data into two subsets, the decision tree (DT) uses the same reasoning to divide each subsequent subset recursively. Until the maximum designated depth is achieved or no further splits that minimize the loss function can be identified, this procedure is repeated[25]. Decision trees (DTs) are widely used classifiers and regressors for constructing binary classification and regression, respectively, because of their simplicity and interpretability. Compared to other algorithms, DTs handle numerical and categorical data efficiently and with fewer assumptions.

RESULTS AND DISCUSSION

Statistical Summary of Water Quality Parameters

A comprehensive statistical analysis of the groundwater quality from two blocks of Begusarai District-Matihani and Cheria Bariarpur-was performed to understand the distribution of key physicochemical parameters and heavy metals. Table 1 presents the water quality statistics for Matihani Block (N=50). pH of samples were found (6.7 to 7.9), with a mean value of 7.3 and a standard deviation (SD) of 0.3142, which lies well within the acceptable range (6.5–8.5) specified by the Indian Standard (IS:10500, 2012). However, elevated concentrations of arsenic (As) were observed, ranging from 0.0163 mg/L to 0.203 mg/L, with a mean of 0.129 mg/L. This mean concentration significantly exceeds the permissible limit of 0.01 mg/L, suggesting a potential public health risk due to long-term exposure. Iron (Fe) levels in the water ranged from 0.0031 mg/L to 0.398 mg/L, with an average concentration of 0.0412 mg/L, which remains within the acceptable limit of 0.3 mg/L.

Table 1: Statistical description of Matihani Block, Begusarai District, N=50

Parameters	Min.	Max.	Mean	S.D.	Acceptable limit IS:10500,2012
Physicochemical Parameter					
pH	6.7	7.9	7.3	0.3142	6.5-8.5
Heavy metals (mg/L)					
As	0.0163	0.203	0.1290	0.0433	0.01
Fe	0.0031	0.398	0.0412	0.0632	0.3

Table 2: Statistical description of Cheria Bariarpur block, Begusarai District, N=36

Parameters	Min.	Max.	Mean	S.D.	Acceptable limit IS:10500, 2012
Physicochemical Parameters					
pH	6.5	7.9	7.275	0.297	6.5-8.5
Heavy metals (mg/L)					
As	0.0007	0.5900	0.1656	0.1299	0.01
Fe	0.0020	0.2080	0.0600	0.0550	0.30

Similarly, the groundwater quality data for Cheria Bariarpur Block (N=36) is summarized in Table 2. The pH values ranged from 6.5 to 7.9, with a mean of 7.275 and a SD of 0.297, also indicating a neutral to slightly basic nature of the water and compliance with the IS standard. Average arsenic concentration was found to be 0.1656 mg/L, with values as high as 0.5900 mg/L, far exceeding the acceptable limit. This block exhibits an even higher mean arsenic level than Matihani, highlighting a more severe contamination issue. The mean concentration of iron was 0.0600 mg/L, varying between 0.0020 mg/L and 0.2080 mg/L, which, like Matihani, remained within permissible limits.

These results indicate that while the general physicochemical characteristics of groundwater (such as pH) are within acceptable bounds, the widespread occurrence of arsenic contamination in both blocks poses a significant environmental and health concern. The spatial variability in arsenic concentrations suggests possible geogenic sources, such as the dissolution of arsenic-bearing minerals, although anthropogenic contributions cannot be ruled out. Such findings underscore the necessity for regular monitoring and the implementation of suitable mitigation strategies to ensure safe drinking water for the affected population.

Evaluation of Machine Learning Model Performance

To predict groundwater quality and associated risk parameters with improved accuracy, multiple machine learning models were employed, and their performance was evaluated based on standard statistical indices, namely Mean Squared Error, Root Mean Squared Error, Nash–Sutcliffe Efficiency, and the coefficient of determination (R^2). Table 3 summarizes the comparative performance of the Decision Tree Regressor and the Random Forest Regressor.

Table 3: Performance of various ML models

Models	Performance criteria (Accuracy for testing sets)			
	MSE	RMSE	NSE	R^2
DTRM	0.002356	0.048538	0.874	0.874
RFRM	0.001867	0.043208	0.895	0.895

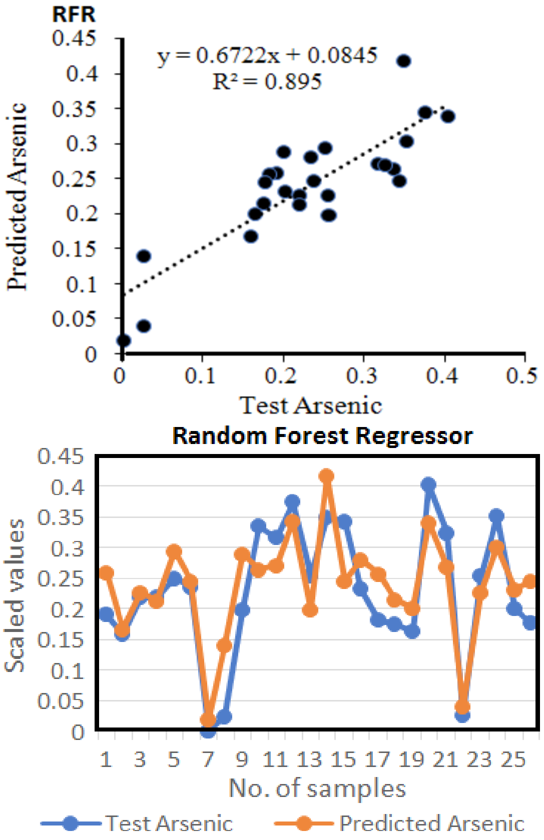

Fig. 3. Prediction accuracy plots for RFR

The DTR exhibited a MSE of 0.002356 and an RMSE of 0.048538, with both NSE and R^2 values standing at 0.874. These metrics indicate that DTRM can capture nonlinear relationships in the dataset reasonably well. However, the RFRM outperformed the DTRM, achieving a lower MSE (0.001867) and RMSE (0.043208), along with higher NSE and R^2 values of 0.895. The superior performance of RFRM can be attributed to its ensemble-based nature, which reduces overfitting and enhances generalization by aggregating predictions from multiple decision trees.

Overall, the results suggest that Random Forest is a more reliable and robust model for predicting groundwater quality parameters in arsenic-affected regions. Its improved accuracy and predictive efficiency make it a valuable tool for water quality assessment and management. Integrating such data-driven approaches with field-level water testing can significantly improve decision-making processes for public health and resource planning.

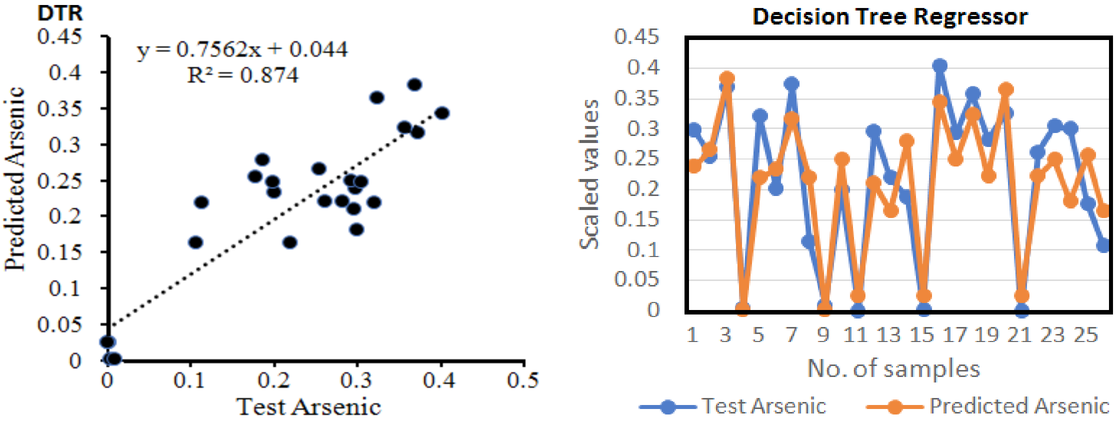

Fig. 4. Prediction accuracy plots for DTR

CONCLUSION

This study presents a comprehensive assessment of groundwater quality in two blocks of Begusarai District, Bihar-Matihani and Cheria Bariarpur-focusing on key physicochemical parameters and toxic heavy metals. While pH levels in both regions remain within the acceptable range defined by IS: 10500 (2012), arsenic concentrations significantly exceed permissible limits, posing serious health risks. Iron levels were found to be within safe boundaries, although variations were observed across locations. The elevated arsenic contamination suggests a geogenic origin, but the possibility of anthropogenic contributions cannot be dismissed. To enhance groundwater quality prediction, machine learning models were employed and evaluated. Among the tested models, the Random Forest Regression Model demonstrated superior performance with lower error values and higher predictive efficiency compared to the Decision Tree Regression Model. These findings highlight the potential of ensemble learning techniques in supporting water quality monitoring and risk assessment frameworks. In conclusion, the combination of statistical analysis and machine learning provides a reliable methodology for identifying contamination hotspots and forecasting groundwater quality. The study emphasizes the urgent need for targeted mitigation strategies and continuous monitoring, especially in arsenic-prone areas, to ensure the provision of safe drinking water and protect public health.

ACKNOWLEDGEMENT

We want to thank the lab facility of NIT Patna and also like to mention that no funding was obtained for this study.

Conflict of interest

The authors declare that we have no conflict of interest.

REFERENCES

1. WHO., Guidelines for Drinking-Water Quality SECOND ADDENDUM TO THIRD EDITION WHO Library Cataloguing-in-Publication Data., World Health Organization., **2008**, *1*, 1–103.http://www.who.int/water_sanitation_health/dwq/secondaddendum20081119.pdf.

2. WHO., Safe Water Technology for Arsenic Removal. UNICEF: **2001**, 1–22.
3. Kanth, K.M., Singh, S.K., Kashyap, A., Vijay Kumar Gupta, V.K., Shalini, S., Kumari, S., Kumari, R., and Puja, K. Bacteriological assessment of drinking water supplied inside the Government schools of Patna District, Bihar, India., *Am. J. Environ. Prot.*, **2018**, *6*(1), 10-13. http://pubs.sciepub.com/env/6/1/2/index.html.
4. Sarath Prasanth, S.V.; Magesh, N.S.; Jitheshlal, K.V.; Chandrasekar, N. and Gangadhar, K.J.A.W.S., Evaluation of groundwater quality and its suitability for drinking and agricultural use in the coastal stretch of Alappuzha District, Kerala, India., *Appl. Water Sci.*, **2012**, *2*, 165-175.
5. Sappa, G.; Ergul, S.; Ferranti, F.; Sweya, L.N. and Luciani, G. Effects of seasonal change and seawater intrusion on water quality for drinking and irrigation purposes, in coastal aquifers of Dar es Salaam, Tanzania., *J. Afr. Earth Sci.*, **2015**, *105*, 64-84.http://dx.doi.org/10.1016/j.jafrearsci.2015.02.007.
6. Mishra, A. K., & Maurya, N. S. Assessment of Groundwater Quality of Patna Urban and Sub-Urban Areas For Its Uses as Drinking and Irrigation Water., *Environ. Eng. Manag. J.*, **2024**, *23*(8).
7. Stigter, T.Y.; Van Ooijen, S.P.J.; Post, V.E.A.; Appelo, C.A.J. and Dill, A.C. A hydrogeological and hydrochemical explanation of the groundwater composition under irrigated land in a Mediterranean environment, Algarve, Portugal., *J. Hydrol.,* **1998**, *208*(3-4), 262-279.
8. Peterson, E.W.; Davis, R.K.; Brahana, J.V. and Orndorff, H.A., Movement of nitrate through regolith covered karst terrane, northwest Arkansas. *J. Hydrol.*, **2002**, *256*(1-2), 35-47.
9. Reghunath, R., Murthy, T.S. and Raghavan, B.R., 2002. The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. *Water Res.,* **2002**, *36*(10), 2437-2442.
10. Basu, A.; Saha, D.; Saha, R.; Ghosh, T. and Saha, B. A review on sources, toxicity and remediation technologies for removing arsenic from drinking water., *Res. Chem. Intermed.*, **2014**, *40*, 447-485.
11. Abeer, N.; Khan, S.A.; Muhammad, S.; Rasool, A. and Ahmad, I. Health risk assessment and provenance of arsenic and heavy metal in drinking water in Islamabad, Pakistan., *Environ. Technol. Innov.*, **2020**, *20*, 101171. https://doi.org/10.1016/j.eti.2020.101171.
12. Jain, C.K. and Ali, I. Arsenic: occurrence, toxicity and speciation techniques., *Water Res.*, **2000**, *34*(17), 4304-4312.
13. Sharma, V.K. and Sohn, M. Aquatic arsenic: toxicity, speciation, transformations, and remediation., *Environ. Int.*, **2009**, *35*(4), 743-759. http://dx.doi.org/10.1016/j.envint. 2009. 01.005.
14. Choong, T. S.; Chuah, T.G.; Robiah, Y.; Koay, F.G. and Azni, I., Arsenic toxicity, health hazards and removal techniques from water: an overview., *Desalination.*, **2007**, *217*(1-3), 139-166.
15. Uddh-Söderberg, T.E.; Gunnarsson, S.J.; Hogmalm, K.J.; Lindegård, M.B.G. and Augustsson, A.L., An assessment of health risks associated with arsenic exposure via consumption of homegrown vegetables near contaminated glassworks., *Sci. Total Environ.*, **2015**, *536*, 189-197.
16. IARC. Monographs on the Evaluation of Carcinogenic Risks to Humans., *IARC Monographs.*, **2004**, *84*, 41–267.
17. IS 10500. **2012**. Indian Standard Drinking Water-Specification (Second Revision)ICS 13.060.20.
18. Hering, J.G., Risk assessment for arsenic in drinking water: limits to achievable risk levels., *J. Hazard. Mater.*, **1996**, *45*(2-3), 175-184.
19. USEPA. Quantitative Risk Assessment Calculations: Sustainable Futures /P2 Framework Manual., *USEPA.*, **2012**, *748*, 1–11.
20. Bhardwaj, V.; Singh, D.S. and Singh, A.K. Water quality of the Chhoti Gandak River using principal component analysis, Ganga Plain, India. *J. Earth Syst. Sci.*, **2010**, *119*, 117-127.
21. Ibrahim, B.; Ewusi, A.; Ahenkorah, I. and Ziggah, Y.Y. Modelling of arsenic concentration in multiple water sources: a comparison of different machine learning methods. Groundw., *Sustain. Dev.*, **2022**, *17*, 100745 https://doi.org/10.1016/j.gsd.2022.100745.
22. Xiangcao, Z.; Su, C.; Xianjun, X.; Ge, W.; Xiao, Z.; Yang, L. and Pan, H., . Employing machine learning to predict the occurrence and spatial variability of high fluoride groundwater in intensively irrigated areas., *Appl. Geochem.*, **2024**, *167*, 106000.https://doi.org/10.1016/j.apgeochem.2024.106000.

23. Jin, Z.; Shang, J.; Zhu, Q.; Ling, C.; Xie, W. and Qiang, B. RFRSF: Employee turnover prediction based on random forests and survival analysis. In Web Information Systems Engineering–WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, **2020**, Proceedings, Part II 21, 503-515. Springer International Publishing.

24. Chakraborty, M.; Sarkar, S.; Mukherjee, A.; Shamsudduha, M.; Ahmed, K.M.; Bhattacharya, A. and Mitra, A. 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.https://doi.org/10.1016/j.scitotenv.2020.141107.

25. Singh, S.K.; Taylor, R.W.; Pradhan, B.; Shirzadi, A. and Pham, B.T. Predicting sustainable arsenic mitigation using machine learning techniques., *Ecotoxicol. Environ. Saf.*, **2022**, *232*, 113271., https://doi.org/10.1016/j.ecoenv.2022.113271.