A Machine Learning Analysis of Groundwater Heavy Metals Contamination

K Sankari
U.G Scholar
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu, India
sankari.velakumaraswamy@gmail.com

Dr. R Subhashini
Professor
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu, India
subhaagopi@gmail.com

Dr. P. Mohana
Scientist
Centre for Remote Sensing and Geoinformatics
Sathyabama Institute of Science and Technology
mohanaperumalvv@gmail.com

Abstract— This project's objective is to classify as safe and unsafe levels of heavy metals using machine learning. In the southern state of Tamil Nadu, samples of groundwater collected from the Arani Taluk were used to generate these levels. An integrated machine-learning framework was developed during this study for the detection of groundwater contamination. In the Arani Taluk of Tamil Nadu, we collected forty-four samples of groundwater, and to characterize and evaluate the water quality, heavy metals were determined. As a result of seasonal variations, the quality of the water was determined. Trace metals such [11] as Mn, Ni, Co, Fe, Cu, Zn, Pb, and Cr were analysed for the pre-monsoon and post-monsoon seasons. The amount of metal present in groundwater was predicted using machine learning algorithm (Decision Tree). Our goal also includes a comparison of the accuracy between the conventional method and the machine learning model-based method. We have also provided visual graphs and images in addition to the above two features in order to support this study. This is to provide a better understanding of this research.

Our main outcome of this study includes creating an efficient machine learning model to estimate the value of contaminants in a given area and a comparison between post-monsoon groundwater metals and pre-monsoon groundwater metals using the conventional and machine-learning method was successfully made.

Keywords— Machine learning, Groundwater contamination, Decision Tree regression method, Heavy metals.

I. Introduction

Groundwater is contaminated when pollutants such as pesticides or gasoline reach the groundwater and make it unsuitable or unsafe for consumption. There is a high chance that the release of chemicals or pollutants into the environment may lead to the contamination of groundwater [6]. Human activity is almost always responsible for the release of pollutants into the environment.

About 20% of the earth's water is groundwater, along with 1% of it being found in rivers, lakes, soil, and the atmosphere. Climate change and freshwater demand have increased significantly due to an increasing human population. The geochemical and mineral content of aquifers, as well as anthropogenic sources incorporated into the weathering process of minerals and rocks which influence the groundwater reliability. This is followed by leaching and runoff as a result [7].

It is not uncommon for metropolitan areas and other urban areas to widespread and systematic growth. This is due to the fact that new towns are resulting in the formation of urban areas. Sewage systems are a basic requirement for any community, but they are lacking in these regions [5]. By consuming metals and being exposed to them in the environment, metals eventually make their way into the human body [10]. These metals reach targeted organs which are the liver, kidney, and brain. It is their ability to modify these systems that determines the fate of metals in the body. When the body accumulates too much metal in the body, it excretes it through the urine and it accumulates in various tissues as well. Metals become toxic when their concentrations reach a certain level. Today, groundwater pollution is one of the most significant environmental problems we face [8]. Due to their tough toxicity even in small applications, heavy metals are a particular concern among the variety of pollutants affecting water properties due to their wide range of effects. It is important to remember that heavy metals can cause a great deal of health problems. These problems can have different signs and symptoms depending on the nature and number of heavy metals consumed [9].

II. Literature Survey

In the work of S. Kumar and J. Pati, there was a groundwater contamination assessment of arsenic that relied on machine learning. As part of this study, water parameters were used as the input variables for distinguishing the samples of water in accordance with threshold limits (WHO). A significant contribution of this study can be attributed to the fact that approaches of machine learning relying on determinable relevant parameters were used as inputs. These parameters maybe potentially be useful in anticipating the amount of arsenic contamination. Here, they used a variety of machine-learning algorithms, such as Optimized Forests, SPAARCs, CS Forests, Reduced Error Pruning Trees, and Random Forests to distinguish samples as safe or unsafe. The accuracy, precision, recall, as well as the FPR of these models have been evaluated by a number of evaluation criteria.[1]

P. V. Mohana and P. M. Velmurugan According to them, their method relies on group classification. As part of the process, hydro chemical signatures are recognized and identified. We can reveal a lot about the fundamental connection between the parameters by employing this technique.[2]

This dispels any prior beliefs about hydro chemical data matrix. This technique allows for the grouping of various entities into different groups or clusters based on the degree of similarity between them in terms of properties and nature, while inducing variation among the groups. In their study, a further correlation analysis was also conducted in order to

A. K. Visvam Devadoss et al. (eds.), *Proceedings of the 6th International Conference on Intelligent Computing (ICIC-6 2023)*, Advances in Computer Science Research 107,
https://doi.org/10.2991/978-94-6463-250-7_9

confirm and support the causes of the variation and to pinpoint the origin and source of the pollutants.

In Mahipal Singh Sankhla's article and Rajeev Kumar's article, they reviewed a study that reported heavy metal levels in groundwater as well as the degree of contamination [3]. In order to prevent potential health risks, they made a series of recommendations. The metals are extracted from natural sources and dealt in the industries where heavy metals flowed into the bodies of water during industrial revolution. In the same way, heavy metals are also leaked into the environment through waste, whether it comes from domestic or agricultural sources as well as from auto drains. They proposed the following list of human activities that release heavy metals into the environment and water;

a) Treating metal ores.

b) Mining

c) The combustion of fossil fuels and oil products.

d) Disposing of industrial effluents.

e) Residential waste disposal.

f) Discarding from auto exhausts.

g) Usage of pesticides containing significant metal salts

In Husam Baalousha's article, they proposed a method for combining vulnerability mapping and spatial analysis. This would make determining much more effective regionalized groundwater quality monitoring system easier. Areas with a strong potential of pollution are identified and given priority for monitoring via vulnerability mapping [4]. The collected data is then interpreted using geo - statistical methods, and the geographic mapping allocation of various data collected is investigated. The effectiveness of the distribution of monitoring sites is reflected in the accuracy of spatial mapping. The methodology was used to assess the Heretaunga basin in New Zealand.

An interpolation variance analysis was utilized to confirm the geographical distribution of spots, and DRASTIC approach was performed to create an area vulnerability map. According to the study's findings, a few high vulnerability regions are not covered by the existing network. This suggests that the amount of monitoring stations and their distribution are inefficient.

III. Existing Work

Below are few problems listed in existing system:

- It needs a good setup to do the processing as processing of data requires higher processing power.
- User cannot change a lot of parameters.
- Its costly process because computation time increases as number of points increases.
- Error margin is directly proportional to the square of distance between 2 relative points.

IV. Proposed Work

We have incorporated the results using machine learning method and packages along with few python packages which are the most rapidly growing technologies today. The output is obtained by using machine learning with the help of QGIS tool/software.

In proposed system, time is comparatively reduced, and we can get point wise value whereas in existing system we can only get a value based on range.

The computational power is less here; therefore, the overall cost is low. Here we showed the difference in concentration of heavy metals in ground-water before and after monsoon (Pre-monsoon and Post-monsoon) and also, we showed the difference between the conventional method and machine learning method which was done using google collaboration. We will also provide assurance to the public and work to prevent the spread of false information. The suggested system also attempts to identify any unexpected processes or events as well as changes in contaminant mobility (if they do occur). The proposed idea is to draw a comparison between conventional methodology and the proposed methodology. Further, the same model can be reused for other similar use cases which involve interpolation.

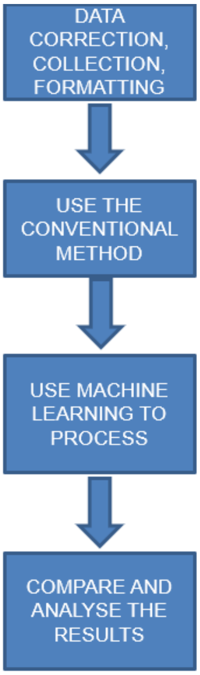
DATA CORRECTION, COLLECTION, FORMATTING
USE THE CONVENTIONAL METHOD
USE MACHINE LEARNING TO PROCESS
COMPARE AND ANALYSE THE RESULTS

Fig 1. Flow chart of process using machine learning

To implement this model, execution of program is done through Google Colab and necessary libraries have to be installed to perform certain functions. Figure 1 shows the steps in the process using machine learning.

In the current study, groundwater samples are collected in a relatively close pattern at sampling stations using a random sampling method. To assess the seasonal impact on groundwater physicochemical parameters, samples were taken at 44 points throughout the study region indicating pre-monsoon and post-monsoon periods. The dataset consists of Latitude, Longitude points and the data points of 8 heavy metal in groundwater namely Zinc, Lead, Manganese, Nickel, Cobalt, Iron, Copper and Chromium each of 44 values.

V. Selected Methodology

For this study, the data has been collected from 44 different collection sites spread across the whole Arani region. Initially, each metal (dataset) is saved in csv format. Then we have extracted the different metal values from csv file and stored it into each metal folder respectively for pre monsoon. Then, we convert the plane shape file into vector shape file based on the requirements and define the study area and store the study area in json format. The desired area is digitized by using ARC GIS software. A boundary enclosing the area is traced manually using an open-source tool. The traced-out boundary coordinates are then exported as a JSON file containing all the border coordinates of the enclosed area. The master data is divided into seven separate datasets based on each metal. Each of the seven subfiles contain four values which are site no, latitude, longitude and concentration level. This separated data is stored as a vector file in a shape file (.shp) format.

A. *Conventional Method:*

We use Gdal (The vector shapefile for each metal is read using Geospatial Data Abstraction Library) to perform interpolation for all the metals. Interpolation is a process of determining the values for unknown points (estimation points) based on the values of known points (data points). There are different types of interpolation algorithms and techniques available. In this study Inverse Distance Weighted (IDW) interpolation technique has been used. Then, we calculated the performance metrices and plotted the data for respective metals. Lastly, we clipped the interpolation result and repeated the entire process for post monsoon metals. In the resultant raster file (Fig. 2, 4, 6, 8, 10, 12, 14, and Fig. 16) the dark colours (Blue shades) show the lowest concentration and the bright colours (Red shades) shows the highest concentration.

B. *Machine Learning Method:*

To implement a machine learning algorithm, the data needs to be processed. In general, for any machine learning model the complete data is split into two sets. The first one is the training set, used to train the Machine Learning (ML) model and the second one is the testing set, used for validating the trained model. These two sets are further split into two different parts. The first one includes all the independent variables i.e., input parameters. The second part is the dependent variable i.e., the output parameter. Usually while splitting the data, the Pareto Principle, also known as 80-20 is followed. Using this split ratio makes sure that at the end there is enough data for both training and testing out the ML model. After interpolating the metals, in order to implement the machine learning method, we use the Decision Tree regression method to train the models.

C. *Decision Tree Regression:*

It is a supervised learning algorithm which works by making an 'N' number of decisions in each iteration. In this algorithm, the whole prediction is carried out by making a regression tree. In this tree, each leaf node has a numeric value which has been predicted. The instances which are created using available modules are fed with the training data to learn some complex linear and non-linear relations between the dependent and independent variables. Once the model instances have been trained, they can be used to predict the results. The trained model will try to fit the learnt relation over the independent variables and the dependent variables supplied from the testing set.

For this study, the latitude and longitude serve as two independent variables while the contamination value of each heavy metal is considered a dependent variable.

VI. Results

Below are the results which shows the difference in concentration during Post-monsoon and Pre-monsoon for the heavy metals using conventional method and machine learning method.

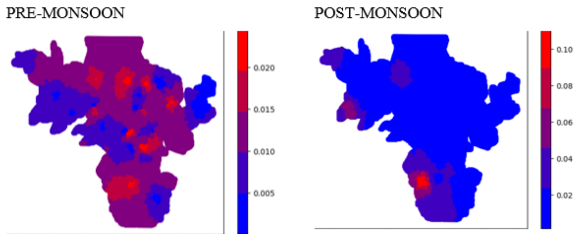

Fig 2. Difference in concentration during Post-monsoon and Pre-monsoon for Cobalt using conventional method.

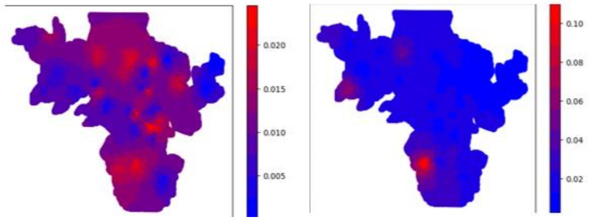

Fig 3. Difference in concentration during Post-monsoon and Pre-monsoon for Cobalt using machine learning method.

We can see in figures 2 and 3 that the concentration of Co metal present is huge during the pre-monsoon and reduces during the post-monsoon i.e., we can see that after rain (post-monsoon) the concentration of Cobalt is drastically reduced.

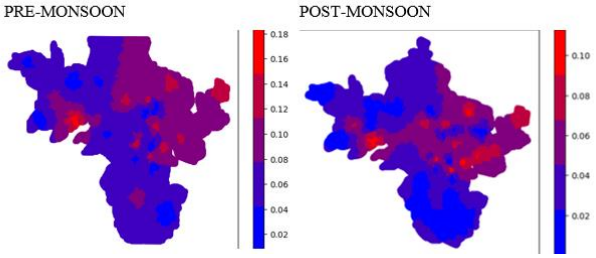

Fig 4. Difference in concentration during Pre-monsoon and Post-monsoon for Chromium using conventional method.

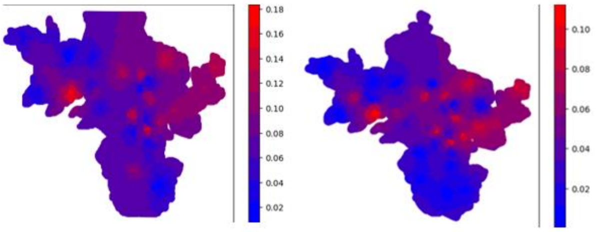

Fig 5. Difference in concentration during Post-monsoon and Pre-monsoon for Chromium using machine learning method.

We can see in figures 4 and 5 that the concentration of Cr metal present is almost same but reduced in some areas during the post-monsoon i.e., we can see that after rain (post-monsoon) the concentration of Chromium is slightly reduced.

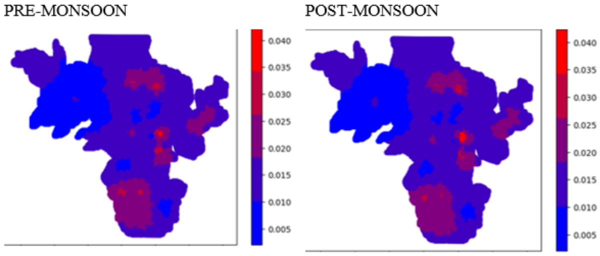

Fig 6. Difference in concentration during Pre-monsoon and Post-monsoon for Copper using conventional method.

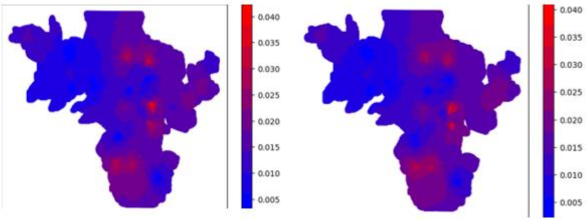

Fig 7. Difference in concentration during Pre-monsoon and Post-monsoon for Copper using machine learning method.

We can see in figures 6 and 7 that the concentration of Cu metal present is almost same during the pre-monsoon and post-monsoon i.e., we can see that even after rain (post-monsoon) the concentration of Copper has no much difference.

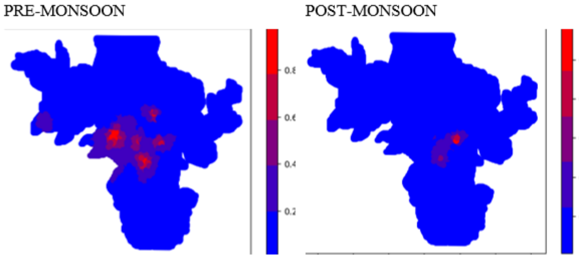

Fig 8. Difference in concentration during Pre-monsoon and Post-monsoon for Iron using conventional method.

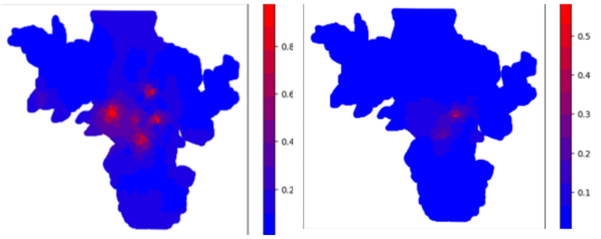

Fig 9. Difference in concentration during Pre-monsoon and Post-monsoon for Iron using machine learning method.

We can see in figures 8 and 9 that the concentration of Fe metal present is more during the pre-monsoon and reduces during the post-monsoon i.e., we can see that after rain (post-monsoon) the concentration of Iron is drastically reduced.

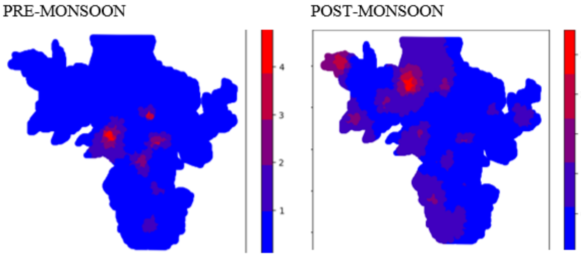

Fig 10. Difference in concentration during Pre-monsoon and Post-monsoon for Manganese using conventional method.

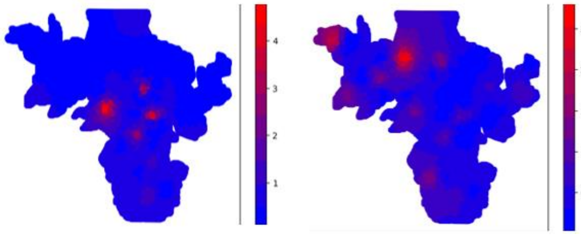

Fig 11. Difference in concentration during Pre-monsoon and Post-monsoon for Manganese using machine learning method.

We can see in figures 10 and 11 that the concentration of Mn metal present is more during the post-monsoon and reduces during the pre-monsoon i.e., we can see that after rain (post-monsoon) the concentration of Manganese is drastically increased and decreased in some places.

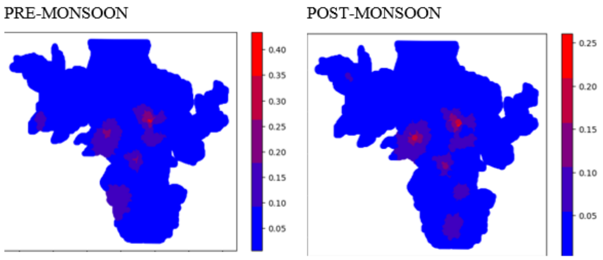

Fig 12. Difference in concentration during Pre-monsoon and Post-monsoon for Nickel using conventional method.

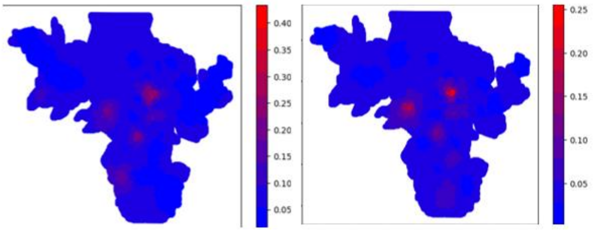

Fig 13. Difference in concentration during Pre-monsoon and Post-monsoon for Nickel using machine learning method.

We can see in figures 12 and 13 that the concentration of Ni metal present is almost same during the pre-monsoon and post-monsoon i.e., we can see that even after rain (post-monsoon) the concentration of Nickel is almost same.

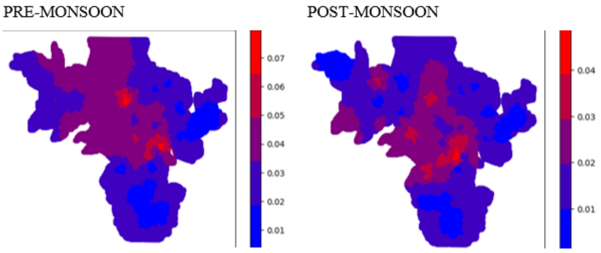

Fig 14. Difference in concentration during Pre-monsoon and Post-monsoon for Lead using conventional method.

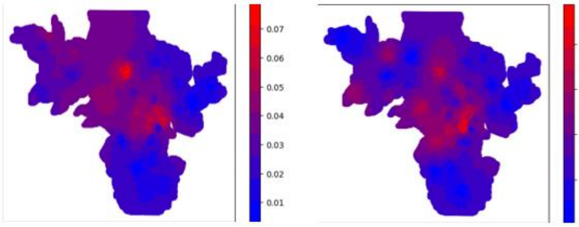

Fig 15. Difference in concentration during Pre-monsoon and Post-monsoon for Lead using machine learning method.

We can see in figures 14 and 15 that the concentration of Pb metal present is more during the pre-monsoon and reduces during the post-monsoon i.e., we can see that after rain (post-monsoon) the concentration of Lead is almost same and is minutely increased infew areas.

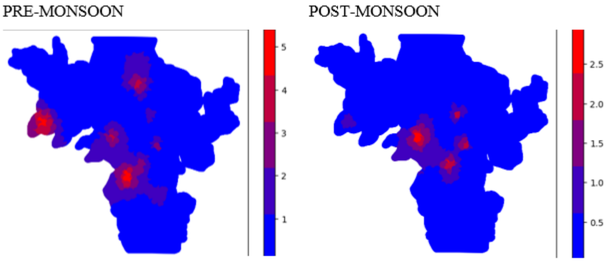

Fig 16. Difference in concentration during Post-monsoon and Pre-monsoon for Zinc using conventional method.

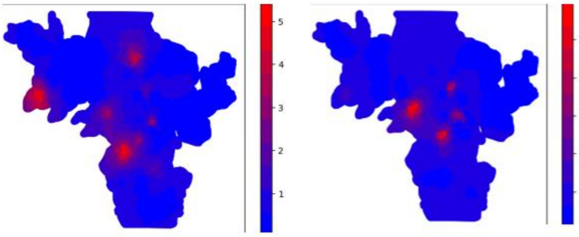

Fig 17. Difference in concentration during Post-monsoon and Pre-monsoon for Zinc using machine learning method.

We can see in figures 16 and 17 that the concentration of Zn metal present is more during the pre-monsoon and reduces during the post-monsoon i.e., we can see that after rain (post-monsoon) the concentration of Zinc is slightly reduced as well as incresed.

We have also marked the heavy metals as safe or unsafe using the permissible limit/ recommended maximum concentration determined by IWQ (Irrigation Water Quality).

METALS	RECOMMENDED MAXIMUM CONCENTRATION (MG/L)	THE VALUE PREDICTED*	SAFE/ UNSAFE
Cobalt	0.05	0.025	SAFE
Chromium	0.10	0.18	UNSAFE
Copper	0.20	0.040	SAFE
Iron	5.0	0.10	SAFE
Lead	5.0	0.08	SAFE

Manganese	0.20	5	UNSAFE
Nickel	0.20	0.40	UNSAFE
Zinc	2.0	5	UNSAFE

TABLE 1: The predicted values using machine learning for pre-monsoon.

METALS	RECOMMENDED MAXIMUM CONCENTRATION (MG/L)	THE VALUE PREDICTED*	SAFE/ UNSAFE
Cobalt	0.05	0.10	UNSAFE
Chromium	0.10	0.10	SAFE
Copper	0.20	0.040	SAFE
Iron	5.0	0.5	SAFE
Lead	5.0	0.05	SAFE
Manganese	0.20	2.5	UNSAFE
Nickel	0.20	0.25	UNSAFE
Zinc	2.0	2.5	UNSAFE

TABLE 2: The predicted values using machine learning for post-monsoon.

VII. Conclusion

The objective of this project is to show that, the prediction of concentration during pre-monsoon and post-monsson is accurate while using Machine Learning algorithm and pretty much not accurate in conventional method. In the Arani Taluk of Tamil Nadu, we collected forty-four samples of groundwater, and to characterize and evaluate the water quality, hence classified it as heavy metals and trace metals. As a result of seasonal variations, the quality of the water was determined. Trace metals such as Mn, Ni, Co, Fe, Cu, Zn, Pb, and Cr were analyzed for the pre-monsoon and post-monsoon seasons. After that, we made a comparison of groundwater metals during pre-monsoon and groundwater metals during post-monsoon which is done using both the conventional and using machine learning method for all the heavy metals.

In the comparison of both the results via plotted graphs, calculated using Machine learning algorithm and conventional method, it was observed that the desired outputs have been received and the results are very much accurate. The Machine learning estimated values cover almost all the values estimated using the conventional method. The model gave an accuracy of 98% and more for all metals.

VIII. References

[1] S. Kumar; J. Pati. Assessment of groundwater arsenic contamination using machine learning in Varanasi, Uttar Pradesh, India (2022).

[2] P. Mohana & P. M. Velmurugan. Evaluation and characterization of groundwater using chemometric and spatial analysis. (2020).

[3] Mahipal Singh Sankhla* and Rajeev Kumar. Contaminant of Heavy Metals in Groundwater & its Toxic Effects on Human Health & Environment (2019).

[4] Baalousha, H. Assessment of a groundwater quality monitoring network using vulnerability mapping and geostatistics: A case study from Heretaunga Plains, New Zealand. Agricultural Water Management. (2010).

[5] APHA. Standard methods for the examination of water and wastewater (19th ed., p. 1467). Washington, DC: American Public Association. (1995).

[6] A. Ashraf, X. Chen and R. Ramamurthy. Modelling of heavy metals contamination in groundwater of southern Punjab, Pakistan. (2021).

[7] Ashwani Kumar Tiwari, Prasoon Kumar Singh, Abhay Kumar Singh, Marina De Maio. Estimation of Heavy Metal Contamination in Groundwater and Development of a Heavy Metal Pollution Index by Using GIS Technique. (2016).

[8] Sad ahamed, Mrinal Kumar Sengupta, Amitava mukherjee, M. Amir Hossain, Bhaskar Das, Bishwajit Nayak, Arup Pal, Subhas Chandra Mukherjee, Shyamapada Pati, Rathindra Nath Dutta, Garga Chatterjee, Adressh Mukherjee, Rishi Srivastava, Dipankar Chakraboti. Arsenic groundwater contamination and its health effects in the state of Uttar Pradesh (UP) in upper and middle Ganga plain, India: A severe danger. (2006).

[9] Sudhakar singha, Srinivas Pasupuleti, Soumya sucharita singha, Suresh kumar. Effectiveness of groundwater heavy metal pollution indices studies by deep-learning. (2020).

[10] Roohul Khan. Occurrence and health risk assessment of arsenic and heavy metals in groundwater of three industrial areas in Delhi, India. (2013).

[11] Murat Öztürk et al., “Trends of trace metal (Mn, Fe, Co, Ni, Cu, Zn, Cd and Pb) distributions at the oxic-anoxic interface and in sulfidic water of the Drammensfjord”, Marine Chemistry , Volume 48, Issues 3–4 , February 1995, Pages 329-342.