International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4,2025
https://theaspd.com/index.php

Performance Evaluation of Ensemble Decision Tree in Identification of Metal Contaminants in Water Reservoirs

V.Kalist[1], S.Poornapushpakala[2*], M.Subramoniam[3], S.Barani[4],J.Merlin Mary Jenitha[5]
[1,2,3,4]Department of ECE, Sathyabama Institute of Science and Technology, Chennai, India
poornapushpakalas@gmail.com
[5]Department of IT, Sathyabama Institute of Science and Technology, Chennai, India

Abstract–*Machine Learning Methods are getting popular day by day and finds its applications in all fields. Researchers are involved in developing new machine learning algorithms by optimizing the various parameters. On the other hand, studies are also in progress in analyzing the performance of the developed machine learning algorithms in the various applications. This paper discusses one such study done by applying decision tree and ensemble decision tree methods to identify the type of metal contaminants in the water reservoirs. The colour histogram features extracted from the images of Lemna minor was grown on waters with different metal dissolvent were used for the study. The performance evaluation done on both the methods concluded that the ensemble method shows better performance than the decision tree method with overall accuracy of 94.5% in the classification of metal contaminants in water.*

Keywords–*Machine Learning, Ensemble technique, Metal Contaminants, Water Reservoir, duckweed.*

INTRODUCTION

The nature of water and pesticides used for farming plays a vital role in determining the quality of the food that is being cultivated. Water bodies are contaminated as a result of urbanization and other human activities [1] and the trend is increasing every year significantly. Water pollution is one of the major reasons for several water borne diseases affecting health of living beings. Some types of this water pollution are visible or sensed by humans, however several other types cannot be identified without the aid of laboratory tests. Estimation of metal contaminants in water is the example of this type. Certain metal contaminants in water are very much harmful to the humans who consumes the water directly.

Therefore, it is necessary to assess the pollutants in water resources and take preventative action to lower them. Water contamination is not just a problem in a particular area but it is a worldwide issue. Thus, it is essential to keep track of water quality in order to protect water resources and do necessary action to recover contaminated ones [2]. Several studies have been carried out to assess metal contaminants in water [3-5]. Each of these studies aimed to provide a simplified approach or develop improved techniques to accurately determine the type of contaminant. Few metal evaluation techniques involve traditional laboratory testing of parameters and their statistical analysis, which is a laborious and time-consuming procedure, since most tests must be carried out in a laboratory and it is essential to preserve the water until the end of the study. Stored water may change its properties over time. Water is analyzed to determine major components such as pH, electrical conductivity, total dissolved solids, and the presence of metal contaminants. Elements such as zinc, iron, arsenic, magnesium, lead, nickel, and copper have been studied in groundwater [6,7]. Research is also being conducted on various water resources for agricultural use [8]. The heavy metal contamination is determined using biological indicators such as Lemna minor, daphnia magna, Sinapis alba [9,10,11]. The change in the properties of the species is used as the measure of the metal contaminants in water. Heavy metal pollution Index (HPI) is used as standard for representing the quality of water [12,13,14]. Water quality analysis using color image processing has helped greatly from the integration of algorithms like Random Vector Functional Link (RVFL) and the group method of data handling (GMDH) model [15]. A color histogram is an effective method of exhibiting the distribution of colors in an image. Red, green, and blue intensities are broken down, and the frequency of intensity of each color in an image is aggregated to create this representation. Ensemble machine learning and deep learning are gaining its popularity in recent years in the field of image processing for various applications [16]. Hence, the objective of this study is to develop a simplistic procedure to estimate the type of metal contaminants in water. This can be done by integrating the image features with machine learning techniques. In this study, the color histogram features extracted from the images of Lemna Minor grown on the water are fed to decision tree algorithm. This output produced by the algorithm is used to estimate the type of metal contaminants in the water. This approach is simplistic and also produce instantaneous results without the need of expertise. The same features are also used to develop an enhanced decision tree algorithm by using ensemble methods. The various steps involved in this study is given in section II.

proposed method

Dataset

The block diagram of the proposed system to identify the metal contaminants using enhanced decision tree algorithm is shown in figure. The first step in the process is to frame the data set required for the study. From the literature studied done, the major metal contaminants identified in the water reservoirs around in India are iron (Fe), lead (Pb), zinc (Zn), cadmium (Cd), copper (Cu), mercury (Hg), chromium (Cr), arsenic (As), nickel (Ni) and manganese (Mn). So, the metals to be identified for the study are narrowed down to Mercury, Copper, Arsenic and Lead. To obtain the image features related to these metals, metals were artificially induced into the fresh water. i.e. Four samples of water in each contains a metal of a specific type was prepared. Lemna Minor was placed on these prepared water samples and the change in texture over the leaves of Lemna minor was captured using a digital camera of resolution 64 megapixel, Lemna minor has the property to absorb metal contaminates in the water on which it is grown. Hence Lemna Minor was chosen for this study. The absorbed metal by the Lemna minor makes a change in the texture of the leaves of this plant. Hence this change is recorded at regular interval over a period of 15 days. This process was repeated and finally a total of 180 samples were available for this study. Since the texture of the leaves may change due to intensity and other various parameters, these samples were augmented to a sample size of 250 images from each kind of metal. So finally, 1000 samples were available for the study. Figure 1 shows the block diagram of the proposed methodology along with few samples of the images of Lemna Minor grown on various metal contaminants.

Feature Extraction uisng Color Histogram

The features from an image aids the model to understand the image better as compared to the raw pixel values. This also makes the model to be more robust to the variations occurred during image acquisition. The analysis of the image is made easier when feature is used to extract the information from an image. In this study, as the variation occur in the over the leaves of Lemna Minor color histogram was chosen to extract the features and analyze the image. Color histogram is the simple and efficient way for representing the distribution of colors in an image. This representation is done by breaking the red, green and blue intensities and counting the frequency of intensity of those colours in an image. The 0-255 range for RGB intensities is divided in to bins. Bins are used to estimate the range of intensities falling into those bins. In this study each channel has a size of 8 bins. The extracted features are then normalised to represent in terms of relative frequencies. The distribution of colour histogram features for each class of metals for pixel values 0, 1 and 2 are shown in figure 2. The Colour histogram is computed as given in equation (1)

$$H(r,g,b) = \sum_{x=1}^{W} \sum_{y=1}^{H} \delta(bin(R(x,y)) = r, bin(G(\mathrm{x,y})) = g, bin\big(B(x,y)\big) = b) \quad (1)$$

Where,

$H(r,g,b)$ – Histogram bin value for the RGB Colour combination (r,g,b)

$R(x,y)$, $G(x,y)$, $B(x,y)$ – red, green, blue channel intensities of a pixel at the location (p,q) in the image

W – Width of the Image

H – Height of the Image

δ – indicator function, returns 1 if the pixel Color falls into the bin(r,g,b) ; returns 0otherwise

$bin\big(R(x,y)\big), bin\big(G(x,y)\big), bin\big(B(x,y)\big)$ – bin indices for the pixel's R,G,B channel values

From the figure it is observed that the distribution of pixels differs that fall in bins gets differed with respect to each class of metals. Hence colour histogram features can be used to analyse and classify the image by the classifier. Decision tree and Ensemble decision tree algorithms are used in this study to analyse and classify the features in accordance to the relevant metal.

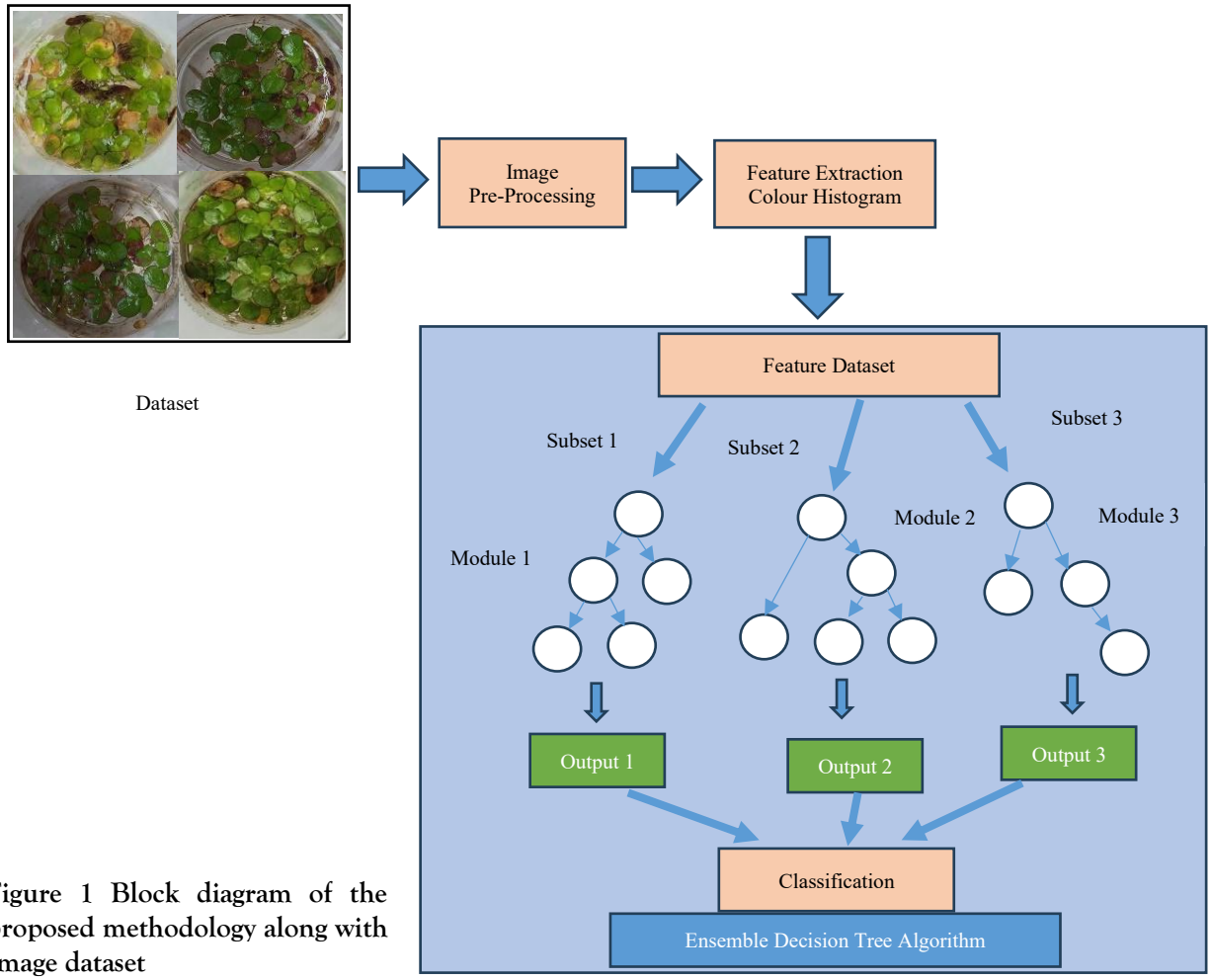
Descicion Tree (DT)Classifier

Decision tree is a supervised learning algorithm which is used to classify the extracted features to its relevant categories. The method of arriving the decision forms a tree like structure. Each node in the tree corresponds to a feature and each branch of the tree denotes a decision rule. Each leaf node is related to an outcome or class label. Gini impurities are used in this study for finding the inequalities in the data and splitting the data for each node. The Gini index can be mathematically expressed as in equation (2)

$$G = 1 - \sum_{i=1}^{n} p_i^2 \quad (2)$$

Where,
G – Gini Index
P_i - proportion of samples belonging to class i at the node
n – Total number of classes

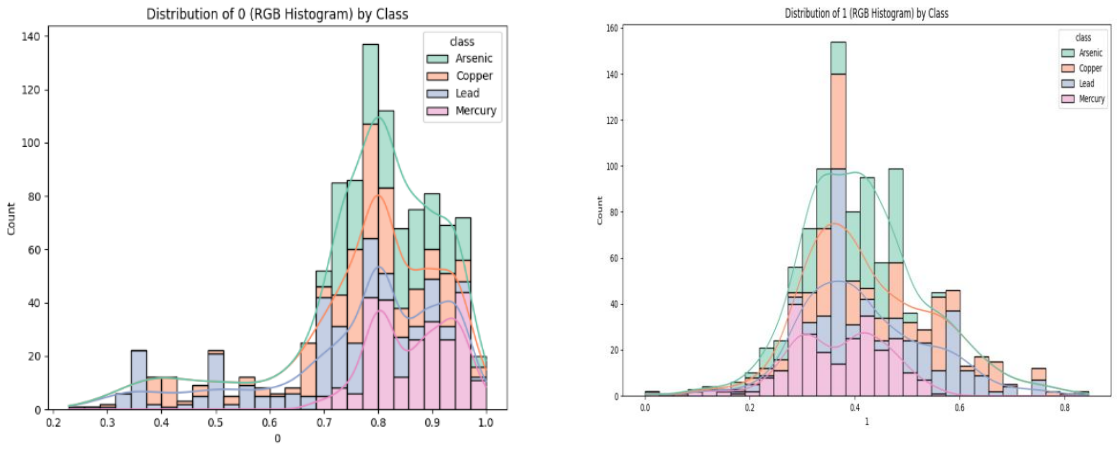
The advantage of the decision tree model is that, the data can be visualized and interpreted without much normalization and standardization. At each node, the algorithm selects the feature and threshold that maximize the reduction in impurity. The color histogram features extracted from 1000 images are fed to this DT classifier. The resultant confusion matrix is shown in figure 3



Dataset

Figure 1 Block diagram of the proposed methodology along with Image dataset

the



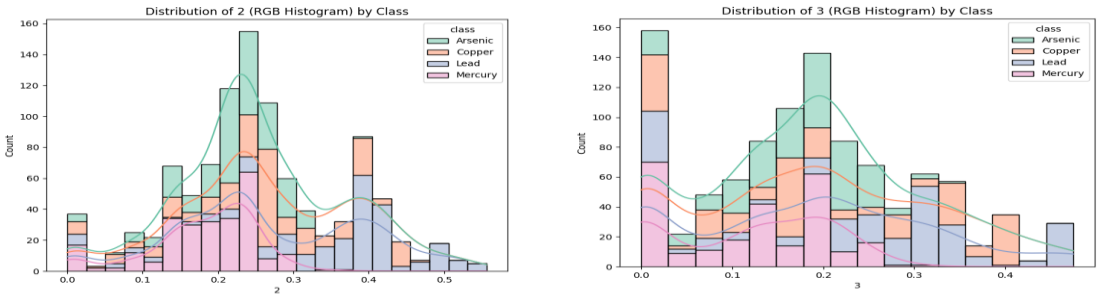


Figure 2 Distribution Color Histogram Features for RGB color intensity values of 0 1 2 and 3 for various class of metal

.

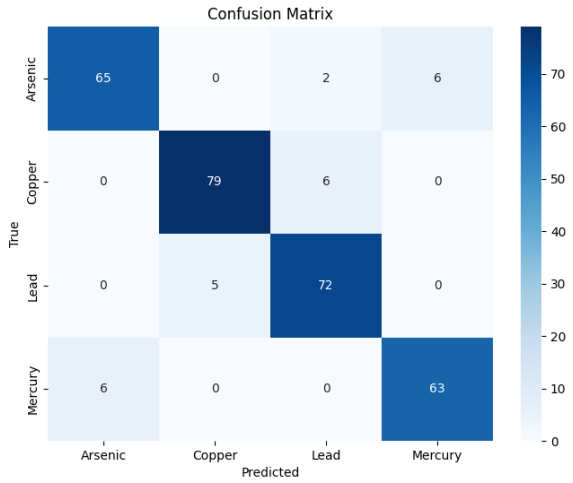


Figure 3 Confusion matrix for Decision Tree model

Table I Performance metrices for DT and Ensemble DT algorithms for metal contaminant classification

Methodology	**Metal contaminant**	**Sensitivity** (%)	**Specificity** (%)	**Overall Accuracy** (%)
Decision Tree	Arsenic	89.04	97.27	91.77
	Copper	92.94	97.56	
	Lead	93.50	96.27	
	Mercury	91.30	97.29	
Ensemble Decision Tree	Arsenic	95.89	97.29	94.07
	Copper	95.29	98.08	
	Lead	94.80	97.70	
	Mercury	89.85	98.67	

Ensemble DT Classifier

Ensemble methods are generally applied where the prediction accuracy is much important or it needs to be improved. This method improves the performance by combining the strengths of several models which makes this method much powerful tool in machine learning. This ensemble methods are less sensitive to noise and overfitting. Based on the nature of data, different algorithms can be combined together to obtain the required solutions. The various types of ensemble methods are bagging, boosting and stacking. In this study bagging method is used. In this method multiple decision tree combined to form the ensembled method. Voting method is used to select the best class from the results produced by individual classification algorithm. This bagging classifier reduces the error by modifying the variance

during the learning process. This results in improved accuracy as compared with the single classifier. The mathematical expression for ensembled bagging classifier is given in equation (3). The confusion matrix obtained with this ensembled decision tree method is given in figure 4.

$$\hat{y} = majority\_vote(f_1(x), f_2(x), f_3(x) \ldots \ldots f_t(x) \quad (3)$$

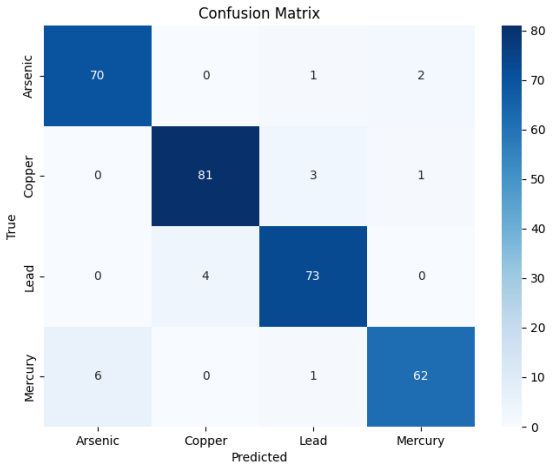


Figure 4 Confusion matrix for ensemble decision tree model

RESULT AND DISCUSSION

The various statistical parameters calculated to evaluate the performance of this combination is given in table I. The comparison chart of the statistical parameters obtained for various class of the metal is given in figure 5 and 6. DT algorithm provided good sensitivity in classifying Lead with 93.5 %. Whereas Ensemble DT algorithm is able to classify Arsenic and Copper with the sensitivity of 95.89 % and 95.29 % respectively. On comparing both the method, the proposed ensemble decision tree shows an improved accuracy of 94.07 % as compared to normal decision tree method with 91.77 % accuracy. Hence, this method can be used for identifying the type of metal contaminants in the water storage systems.

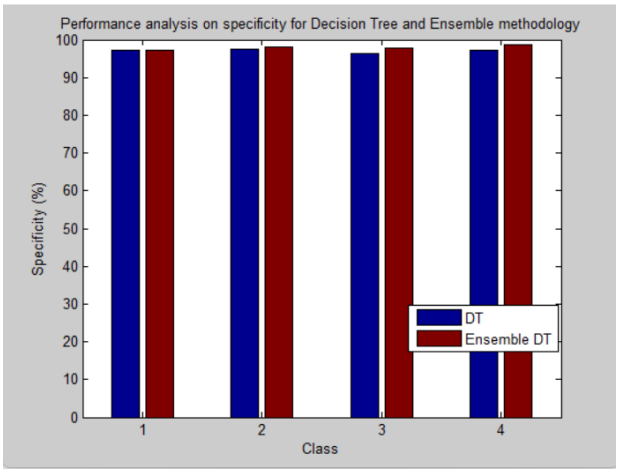


Figure 5 Performance analysis of DT and Ensemble DT algorithms with Sensitivity

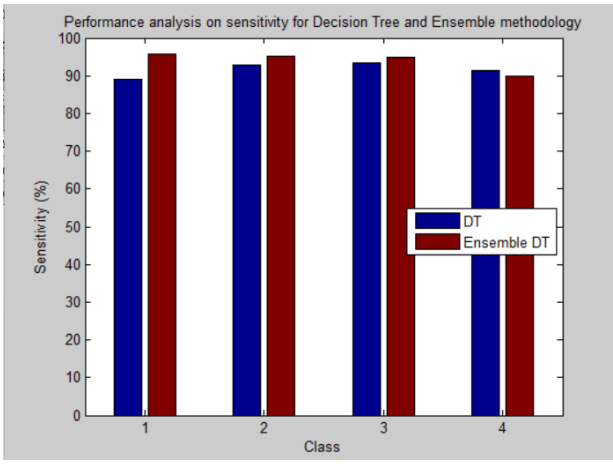


Figure 6 Performance analysis of DT and Ensemble DT algorithms with Specificity

CONCLUSION

Assessment of heavy metal content in water resources is the prime focus of the study. The bioindicator *Lemna minor* is grown on the metal contaminated water. Heavy metals such as Arsenic, Copper, Lead and Mercury are considered for the study. The images of the plant in the metal contaminated water are captured and classified using DT algorithm and ensembled DT algorithm. The Ensemble DT algorithm provided 2.3 % higher accuracy than DT algorithm. This study is done by adding the metal contaminants explicitly. Further studies can be done by growing the plant in natural water. Validation can be done using laboratory studies.

REFERENCES

1. Prachi Vasistha, Rajiv Ganguly, Water quality assessment of natural lakes and its importance: An overview, Proceedings of Materials Today: 32 pp.544–552, 2020
2. Safiur Rahman, M., Shafiuddin Ahmed, A.S., Rahman, M.M. et al. Temporal assessment of heavy metal concentration and surface water quality representing the public health evaluation from the Meghna River estuary, Bangladesh. Appl Water Sci 11, 121, 2021, https://doi.org/10.1007/s13201-021-01455-9
3. D. Rosado, F. Castillo, Nambi, R. Sadhasivam, H. Valleru, N. Fohrer, Evaluating heavy metal levels and their toxicity risks in an urban lake in Chennai, India, International Journal of Environmental Science and Technology 21:1849–1864,2024
4. U. Sai Kiran, Vidhya Lakshmi Sivakumar, An Analysis of Chennai's Lake's Pre- and Post-Monsoon Heavy Metal Pollution, Journal of Survey in Fishery Sciences, 10 (1), 2023, https://doi.org/10.17762/sfs.v10i1S.453
5. M. Tholkappian, R. Ravisankar, A. Chandrasekaran, J. Prince Prakash Jebakumar, K.V. Kanagasabapathy, M.V.R. Prasad, K.K. Satapathy, Assessing heavy metal toxicity in sediments of Chennai Coast of Tamil Nadu using Energy Dispersive X-Ray Fluorescence Spectroscopy (EDXRF) with statistical approach, Toxicology Reports 5 pp. 173-182, ISSN 2214-7500, 2018
6. Zahid Ullah , Abdur Rashid , Junaid Ghani , Javed Nawab , Xian-Chun Zeng, Muddaser Shah , Abdulwahed Fahad Alrefaei , Mohamed Kamel , Lotfi Aleya , Mohamed M. Abdel-Daim and Javed Iqbal, Groundwater contamination through potentially harmful metals and its implications in groundwater management, Frontiers in Environmental Science, 2022, https://doi.org/10.3389/fenvs.2022.1021596
7. Swaminathan, Rajan. (2016). Heavy Metal Contamination in Ground Water of Chennai Metropolitan City, Tamil Nadu, India – A Pilot Study. Biojournal. 11. 13-23.
8. Sajjad Hussain, Ahmad Hassan, Pakiza Arshad, Muhammad Akbar Anjum, Different Sources of Irrigation Water Affect Heavy Metals Accumulation in Soils and Subsequently on Physiological Determinants and Physico-Chemical Properties of Guava Fruits, DOI: https://doi.org/10.21203/rs.3.rs-498294/v1
9. Jihae Park, Eun-Jin Yoo, Kisik Shin, Stephen Depuydt , Wei Li , Klaus-J. Appenroth , Adam D. Lillicrap , Li Xie , Hojun Lee , Geehyoung Kim , Jonas De Saeger , Soyeon Choi , Geonhee Kim , Murray T. Brown and Taejun Han, Interlaboratory Validation of Toxicity Testing Using the Duckweed Lemna minor Root-Regrowth Test, Biology 11, 37, 2022 https://doi.org/10.3390/biology11010037
10. Karel Horak, Jan Klecka, and Miloslav Richter, Water Quality Assessment by Image Processing, Proceedings of International Conference on Telecommunication and Signal Processing, pp.577-581,2015
11. Kazberuk, W.; Szulc, W.; Rutkowska, B. Use Bottom Sediment to Agriculture—Effect on Plant and Heavy Metal Content in Soil. Agronomy 2021, 11, 1077. https://doi.org/ 10.3390/agronomy11061077

12. Priti Saha1 and Biswajit Paul, Assessment of Heavy Metal Pollution in Water Resources and their Impacts: A Review, Journal of Basic and Applied Engineering Research, Volume 3, Issue 8; April-June, 2016, pp. 671-675
13. Y. Sudharshan Reddy, V. Sunitha, Assessment of Heavy metal pollution and its health implications in groundwater for drinking purpose around inactive mines, SW region of Cuddapah Basin, South India, Total Environment Research Themes, Volume 8, 2023,100069, https://doi.org/10.1016/j.totert.2023.100069.
14. Giri, Soma & Singh, Abhay. (2013). Assessment of Surface Water Quality Using Heavy Metal Pollution Index in Subarnarekha River, India. Water Quality, Exposure and Health. 5. 173-182. 10.1007/s12403-013-0106-2.
15. Junde Chen, Defu Zhang, Shuangyuan Yang, Yaser Ahangari Nanehkaran, Intelligent monitoring method of water quality based on image processing and RVFL-GMDH model, IET Image Process 14(17), pp. 4646-4656, 2020
16. Sheik Imran, Pradeep N, A Review on Ensemble Machine and Deep Learning Techniques Used in the Classification of Computed Tomography Medical Images, International Journal of Health Sciences and Research Vol.14; Issue: 1; January 2024, pp. 201- 213