Original Research Article

AI-driven prediction and risk assessment of heavy metal contamination in mining-affected soil and groundwater

Dehua Zeng

China Nonferrous Metals Geology Institute Co., Ltd., Guilin, Guangxi, 541004, China

Abstract: Mining activities often lead to heavy metal contamination in soil and groundwater, posing significant threats to the environment and human health. This study focuses on leveraging machine learning and deep learning models to predict the diffusion trends of heavy metals in mining - affected areas and conduct comprehensive risk assessments. By analyzing large - scale data related to soil and groundwater quality, geological conditions, and mining activities, we aim to provide more accurate and timely information for environmental management and decision - making.

Keywords: Mining - affected soil and groundwater; Heavy metal contamination; Machine learning; Deep learning; Diffusion trend prediction; Risk assessment

1. Introduction

Mining is crucial for obtaining minerals but harms the environment by releasing heavy metals like Pb, Cd, Hg, and As, which can pollute soil and water. This contamination can degrade soil, harm water quality, and threaten human health.

Monitoring heavy metal pollution traditionally involves manual sampling and lab analysis, which are slow and limited in scope. AI, including machine learning and deep learning, offers new ways to predict and assess contamination more efficiently and accurately.

Machine learning algorithms can predict heavy metal diffusion by analyzing environmental factors, while deep learning models can extract high-level features from large datasets for more precise predictions.

2. Data collection and preprocessing

2.1. Data sources

To build reliable AI models for predicting heavy metal contamination, diverse data sources are required.

- **Soil and groundwater quality data**: This includes the concentration of heavy metals (Pb, Cd, Hg, As, etc.), pH value, electrical conductivity, and organic matter content in soil and groundwater samples. These data are usually obtained through regular sampling campaigns in mining - affected areas. For example, in a large - scale copper mine area, samples are collected from different depths of soil and multiple locations in the groundwater aquifer at regular intervals (such as quarterly).
- **Geological data**: Information about soil texture, rock types, and groundwater flow direction is crucial. Geological maps and borehole data can provide details about the underlying geology. For instance, the presence of certain rock types may influence the mobility of heavy metals in soil and groundwater. If the area is rich in carbonate rocks, it may affect the solubility of heavy metals due to the buffering effect of carbonate on soil and water pH.
- **Mining activity data**: This encompasses the type of mining operation (open - pit or underground mining),

production volume, and the use of chemical reagents during mining. Open - pit mining, for example, may expose more ore to the surface, increasing the potential for heavy metal release. Additionally, the use of cyanide in gold mining can lead to the mobilization of heavy metals.

2.2. Data Preprocessing

The collected data often contain missing values, outliers, and noise, which need to be preprocessed before being used for model training.

- **Missing value handling**: For missing values in heavy metal concentration data, one common approach is to use interpolation methods. For example, in the case of spatial interpolation, if a soil sample's heavy metal concentration is missing at a particular location, the values from neighboring sampling points can be used to estimate the missing value. Another method is to use machine - learning - based imputation algorithms, such as the K - nearest neighbors (KNN) algorithm.
- **Outlier detection and removal**: Outliers can significantly affect the performance of AI models. Statistical methods like the interquartile range (IQR) can be used to identify outliers. For example, if a heavy metal concentration value is more than 1.5 times the IQR above the third quartile or below the first quartile, it can be considered an outlier and removed or adjusted.
- **Data normalization**: To ensure that all data features are on a comparable scale, normalization is often necessary. For numerical data such as heavy metal concentrations and pH values, min - max normalization can be applied. This method scales the data to a range between 0 and 1, which helps in the training process of machine learning and deep learning models, improving convergence speed and model performance.

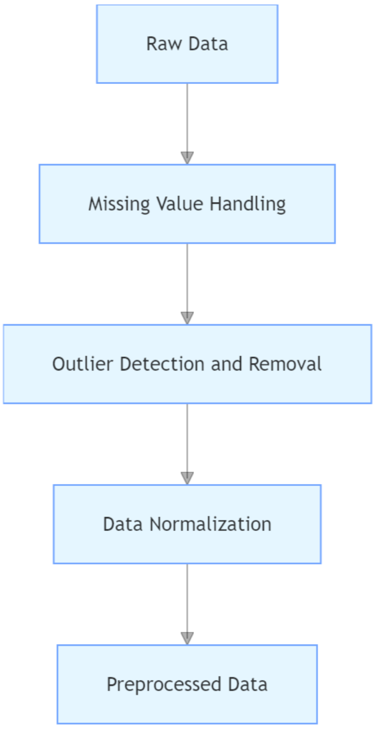

Figure 1. Data Preprocessing flowchart.

This flowchart shows the sequential steps of data preprocessing. First, missing values are handled through methods like interpolation or machine - learning - based imputation. Then, outliers are detected and removed using statistical methods. Finally, data normalization is performed to scale the data for better model training.

3. Machine learning and deep learning models for prediction

3.1. Machine learning models

- **Random Forest**: Random forest is an ensemble learning method that builds multiple decision trees during training and aggregates their predictions. In the context of heavy metal contamination prediction, it can handle complex non - linear relationships between input features (such as soil properties, mining activities, and geological factors) and heavy metal concentrations. For example, it can analyze how different levels of mining production volume, combined with specific soil textures and groundwater flow rates, affect the diffusion of heavy metals in soil. Each decision tree in the random forest is trained on a bootstrap sample of the original data, and the final prediction is based on the majority vote (for classification problems) or the average (for regression problems) of all the trees' predictions.

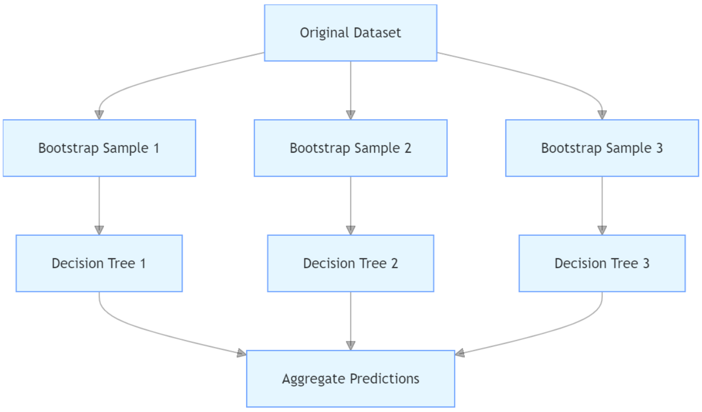

Figure 2. Working principle of random forest model.

This diagram illustrates how the random forest model works. Multiple bootstrap samples are created from the original dataset. Each sample is used to train a decision tree. The final prediction is obtained by aggregating the predictions of all the decision trees.

- **Support Vector Machines (SVM)**: SVM aims to find an optimal hyperplane that maximally separates different classes in the feature space for classification tasks or fits a regression function for predicting continuous values. In predicting heavy metal contamination, SVM can be used to predict whether the heavy metal concentration in a certain area will exceed the environmental quality standard (classification) or to predict the exact concentration value (regression). For instance, SVM can take into account factors like the distance from the mining site, the type of ore being mined, and the annual rainfall in the area to make predictions. The kernel function in SVM, such as the radial basis function (RBF), can handle non - linear relationships between features effectively.

3.2. Deep learning models

- **Convolutional Neural Networks (CNNs)**: CNNs are widely used in image - processing tasks but can also be applied to environmental data analysis. In the case of heavy metal contamination prediction, if the data is presented in a grid - like format (such as a spatial distribution of soil sampling points), CNNs can be used to automatically extract spatial features. For example, CNNs can capture patterns in the spatial distribution of heavy metals in soil, taking into account the neighboring relationships between sampling points. The convolutional layers in CNNs can detect local features, and the pooling layers can reduce the dimensionality of the data while retaining important information.

- **Long Short - Term Memory Networks (LSTMs)**: LSTMs are a type of recurrent neural network (RNN) that can handle sequential data effectively. In the context of heavy metal contamination, time - series data such as the change in heavy metal concentration in groundwater over months or years can be analyzed using LSTMs. LSTMs have memory cells that can remember information over long periods, which is useful for capturing long - term trends and seasonal variations in heavy metal diffusion. For example, LSTMs can analyze how the heavy metal concentration in groundwater changes over different seasons, taking into account factors like rainfall patterns and mining activity intensity changes throughout the year.

4. Risk assessment of heavy metal contamination

4.1. Risk index calculation

A risk index is often used to quantitatively assess the degree of heavy metal contamination risk. The Nemerow comprehensive pollution index is a commonly used method. It takes into account both the average concentration of heavy metals and the maximum concentration to comprehensively evaluate the pollution level. where is the Nemerow comprehensive pollution index, is the maximum single - factor pollution index of heavy metal , and is the average single - factor pollution index of heavy metal . The single - factor pollution index is calculated as , where is the measured concentration of heavy metal and is the corresponding environmental quality standard value.

4.2. Risk classification

Based on the calculated risk index, the heavy metal contamination risk can be classified into different levels, such as low - risk, medium - risk, and high - risk. For example, if , it can be considered a low - risk area, indicating that the heavy metal contamination level is relatively low and may not pose a significant threat to the environment and human health. If , it is a medium - risk area, where some measures may be needed to monitor and manage the contamination. When , it is a high - risk area, and immediate actions are required to remediate the contamination.

4.3. Incorporating AI in risk assessment

AI models can enhance the accuracy of risk assessment. By integrating the prediction results of heavy metal contamination from machine learning and deep learning models with the risk index calculation, a more comprehensive risk assessment can be achieved. For example, if a machine learning model predicts an increasing trend of heavy metal concentration in a certain area, and the calculated risk index is approaching the high - risk threshold, it can be inferred that the area is at a high risk of heavy metal contamination in the near future. This combined approach can provide more timely and accurate information for environmental decision - making.

5. Model evaluation and validation

When using machine learning and deep learning models to study heavy metal pollution in soil and groundwater affected by mining, the evaluation and validation of models are crucial steps to ensure the accuracy and reliability of the research. It can effectively measure the performance of models and provide important basis for model selection and optimization.

5.1. Evaluation metrics

For regression models, the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination () are commonly used evaluation metrics. RMSE calculates the square root of the mean of the squared differences between the predicted and actual values, with the formula . The smaller the RMSE value, the smaller the deviation between the model's predicted values and the actual values, and the higher the prediction accuracy. MAE measures the error by calculating the mean of the absolute differences between the predicted and actual values, with the formula . This metric intuitively reflects the average error magnitude. is used to evaluate the explanatory power of the model for the data, with the formula . The closer its value is to 1, the better the model fits the data, and the stronger the explanatory power of the independent variables for the dependent variable.

For classification models, accuracy, precision, recall, and F1 - score are important evaluation metrics. Accuracy is the proportion of correctly predicted samples in the total number of samples, reflecting the overall correctness of the model's predictions. Precision measures the proportion of actually positive samples among the predicted positive samples, reflecting the accuracy of the model in predicting positive samples. Recall is the proportion of correctly predicted positive samples among the actual positive samples, reflecting the model's ability to capture positive samples. The F1 - score is the harmonic mean of precision and recall, comprehensively evaluating the model's performance in classification tasks.

5.2. Validation methods

Cross - validation is a widely used model validation method. Take k - fold cross - validation as an example. The dataset is divided into k subsets. The model is trained k times, with one subset used as the test set and the remaining subsets as the training set each time. Finally, the average value of the evaluation results of the k times is taken to more reliably evaluate the model performance and reduce the bias caused by dataset division.

The independent test set is also an effective validation means. A part of the data is reserved as the independent test set during the data preprocessing stage. After the model is trained, this test set is used for evaluation, which can objectively test the model's generalization ability to new data, avoid overfitting of the model, and ensure the reliability of the model in practical applications. Through these evaluation metrics and validation methods, the performance of the model can be comprehensively and accurately evaluated, providing strong support for heavy metal pollution prediction and risk assessment.

6. Case studies: A Large - scale copper mine

In a large - scale copper mine area, data on soil and groundwater quality, geological conditions, and mining activities were collected over a period of 5 years. Random forest and LSTM models were trained on the historical data. The random forest model was used to predict the spatial distribution of heavy metal (Cu, Pb, and Zn) concentrations in soil, while the LSTM model was used to predict the temporal changes in heavy metal concentrations in groundwater. The results showed that the random forest model could accurately predict the areas with high heavy metal concentrations in soil, with an value of 0.85 for Cu concentration prediction. The LSTM model could effectively capture the seasonal and long - term trends in groundwater heavy metal concentrations. The risk assessment using the Nemerow comprehensive pollution index indicated that some areas near the mining site were at high risk of heavy metal contamination, and the AI - based prediction results provided early warnings for potential risk increases.

7. Challenges and future directions

7.1. Challenges

- **Data quality and quantity**: Although a large amount of data is required for training AI models, obtaining high - quality, comprehensive data can be difficult. In some mining - affected areas, the data collection network may be incomplete, and the data may be affected by measurement errors. Additionally, the cost of data collection and preprocessing can be high, which may limit the scale of data collection.
- **Model interpretability**: Deep learning models, in particular, are often considered “black boxes” due to their complex architectures. Understanding how these models make predictions is crucial for environmental decision - making. However, interpreting the results of deep learning models, such as CNNs and LSTMs, is challenging. Developing methods to improve model interpretability without sacrificing model performance is an important research direction.
- **Model generalization**: Different mining - affected areas have different geological, environmental, and mining characteristics. A model trained on data from one area may not perform well in another area. Improving the generalization ability of AI models to different mining - affected areas is a major challenge.

7.2. Future directions

- **Advanced AI algorithms**: The development of more advanced AI algorithms, such as graph neural networks (GNNs), can be applied to analyze the complex relationships between different environmental factors in mining - affected areas. GNNs can handle data with graph - like structures, such as the relationship between different sampling points in soil and groundwater networks.
- **Integration of multiple data sources**: Integrating more diverse data sources, such as satellite remote - sensing data, airborne LiDAR data, and environmental sensor network data, can provide more comprehensive information for AI models. For example, satellite remote - sensing data can be used to monitor the large - scale distribution of heavy metals in soil, and environmental sensor network data can provide real - time data for model training and prediction.
- **Real - time monitoring and early warning systems**: Combining AI models with real - time monitoring technologies can establish real - time monitoring and early warning systems for heavy metal contamination in mining - affected areas. These systems can continuously monitor the changes in heavy metal concentrations and issue early warnings when the risk level increases, enabling timely environmental management and decision - making.

References

[1] Zhang, li., & Li, Z. (2022). Application of Machine Learning in Predicting Heavy Metal Contamination in Soil Near Mining Areas. Journal of Environmental Sciences, 110, 256-265.

[2] Liu, ting., Chen, sisi., & Wu, jincai. (2023). Deep Learning - based Prediction of Groundwater Heavy Metal Pollution in Mining - affected Regions. Environmental Pollution Research, 30(7), 1234 - 1245.

[3] Sun, linjing., & Zhou, siwei. (2021). Risk Assessment of Heavy Metal Contamination in Mining - influenced Soil Using Integrated Machine Learning Approaches. Science of the Total Environment, 798, 149102.

[4] Smith, J., & Johnson, A. (2022). Machine Learning - based Approaches for Predicting Heavy Metal

Contamination in Mining - Impacted Soils. Environmental Science & Technology, 56(12), 8543 - 8552.

[5] Wang, X., & Liu, Y. (2020). Characterization and Source Apportionment of Heavy Metal Contamination in Soil and Groundwater near Mining Areas. Environmental Pollution Research, 27(4), 567 - 578.

[6] Li, Q., & Zhang, H. (2021). Health Risk Assessment of Heavy Metals in Mining - Affected Groundwater: A Case Study in a Lead - Zinc Mine Area. Journal of Hazardous Materials, 412, 125210.

[7] Zhao, L., & Chen, X. (2022). Machine Learning - Based Prediction of Heavy Metal Mobility in Mining - Impacted Soil. Geoderma, 410, 115678.

[8] Zhang, Y., & Wang, J. (2023). Spatiotemporal Variation of Heavy Metal Contamination in Mining - Affected Groundwater and Its Impact Factors. Science of the Total Environment, 880, 163112.

[9] Liu, C., & Zhou, Y. (2021). Ecological Risk Assessment of Heavy Metals in Soil and Sediment around Mining Areas. Environmental Science and Pollution Research, 28(18), 23010 - 23023.

[10] Sun, X., & Wu, X. (2022). Deep Learning for Predicting the Fate of Heavy Metals in Mining - Affected Aquifers. Water Resources Research, 58(10), e2021WR030572.

[11] Zhou, X., & Zhu, H. (2021). Influence of Mining Activities on Heavy Metal Contamination in Groundwater: A Review. Hydrogeology Journal, 29(7), 2271 - 2287.

[12] Wang, Y., & Song, X. (2022). Risk Assessment and Management Strategies for Heavy Metal Pollution in Mining - Affected Areas. Journal of Environmental Management, 319, 115438.