Kufa Journal of Engineering
Vol. 17, No. 2, April 2026, P.P. 1-16
Article history: Received 11 November 2024, last revised 20 December 2024, accepted 21 January 2025

ADVANCED MACHINE LEARNING MODELS FOR PREDICTING DIFFUSION OF POLLUTION IN SOILS

Shaymaa Alsamia[1,2], Edina Koch[1], and Katalin Bene[3]

[1] Department of Structural and Geotechnical Engineering, Széchenyi István University, Hungary.

[2] Faculty of Engineering, University of Kufa, Iraq, Email: shaymaam.alsameea@uokufa.edu.iq.

[3] Department of Transport Infrastructure and Water Resources Engineering, Faculty of Architecture, Civil and Transportation Engineering, Széchenyi István University, Hungary.

https://doi.org/10.30572/2018/KJE/170201

ABSTRACT

the infiltration of hazardous chemicals into the soil causes soil pollution which poses significant risks to ecosystems and human health. For this reason, accurate predicting the diffusion of pollution in soils is important and critical for monitoring and protection the environmental state. In this study we have compared advanced machine learning ML models to predict vertical and horizontal pollution diffusion using complex and multimodal soil experimental datasets. Support vector regression, linear regression, gradient boosting regression, xgboost regression, k-nearest neighbours, and artificial neural networks were employed to build predicted models and compared with each other. The comparison criteria are measuring mean squared error, root mean squared error, mean absolute error, and R-squared as the metrics used to evaluate the predictive models performance. The observed results demonstrate that ensemble methods XGBoost and random forest, outperform other models in predicting pollution diffusion while XGBoost achieving the highest accuracy. On the other hand, linear regression was the least effective while k-nearest neighbours and artificial neural networks showed moderate performance.

KEYWORDS

Pollution Diffusion; Soil Contamination; Machine Learning Models; Environmental Modeling; Predictive Analytics.

1. INTRODUCTION

Soil contamination is a critical environmental issue that poses significant risks to human health and ecosystems (S. M. Alsamia, Mahmood, and Akhtarpour 2020). Contamination occurs when hazardous substances such as heavy metals, pesticides, industrial chemicals, and petroleum by-products, and infiltrate, leachate, etc., infiltrate the soil (Salim, Mohammed, and Fattah 2022; Amiri et al. 2022). The pollutant materials may stay for long time times disrupting soil microbial ecosystems and lead to problematic issues of soil fertility (S. Alsamia, Albedran, and Mahmood 2022).Also, the diffusion of contaminants into the soil can contaminate groundwater and that is affecting possible drinking and agricultural water supplies. In this context, accurate modeling and predicting the diffusion behavior of pollutants in soil has a great deal of importance to environmental monitoring and risk assessment. Studies were proposed physical and empirical models predict pollution diffusion (Xue et al. 2023; Samborska-Goik and Pogrzeba 2024) and these methods often struggle to capture real-world data complexity and nonlinear interactions (S. Alsamia, Koch, and Hamadi 2023; Luo et al. 2023). As a result, machine learning models (Xiang et al. 2024; S. Alsamia and Koch 2024; Bosu et al. 2023) considered as a powerful tools for handling such complexities and providing precise predictions (Haggerty et al. 2023; S. Alsamia and Koch 2023). Recent advances in ML techniques are able to model nonlinear, high-dimensional, and multimodal datasets (Liu et al. 2022). By utilizing machine learning, researchers and technicians would learn more about contaminants' diffusion processes (Ye et al. 2020), thus improving predictive accuracy and offering better decision-making support for environmental remediation efforts (Obead, Omran, and Fattah 2021; Al-Ani, Fattah, and Al-Lamy 2009). On the other hand, optimization algorithms (Albedran and Jármai 2023; Li et al. 2022; Jalghaf et al. 2023) are used extensively to enhance machine learning models (S. Alsamia, Albedran, and Jármai 2022),(Hazim Nasir Ghafil, László, and Jármai 2019) in wide ranges of applications like robotics (Tahmasebi et al. 2020; Hazim Nasir Ghafil and Jármai 2020; Mayer, Szilágyi, and Gróf 2020), and optimal design (Maghawry et al. 2021; H.N. Ghafil and Jármai 2019; Martins and Ning 2021). Also, Control systems have been advanced by the power of metaheuristics (Albedran, Alsamia, and Koch 2025; Albedran, 2025). Accurately predicting the diffusion of pollution in soils remains a significant challenge due to the inherent complexity of soil systems and the nonlinear interactions between multiple environmental factors and traditional models often fail to capture the multimodal relationships present in real-world datasets. This limitation hampers their predictive accuracy and limits their applicability in environmental monitoring. Moreover, existing research on machine learning for

environmental modeling is often constrained by limited datasets, oversimplified assumptions, or a narrow focus on specific algorithms.

The current study utilize advanced machine learning techniques to address the limitations of traditional modeling approaches for predicting pollution diffusion in soils. The goal is to explore and compare the effectiveness of various machine learning models in predicting the spread diffusion of pollution in soils. Support vector regression SVR, gradient boosting regression GBR, XGBoost Regression, artificial neural networks ANN, random forest regression RFR, Linear Regression LR, and K-nearest neighbours KNN were evaluated on a multimodal soil contamination dataset based on performance metrics such as MSE, R-squared (R^2), MAE, and RMSE. As finding, the ensemble learning methods, particularly XGBoost and random forest, outperformed other approaches in predictive accuracy.

The key contributions of this study are as follows:

- This research introduces and utilizes a unique, multimodal soil contamination dataset, which captures both vertical and horizontal pollution diffusion, providing a comprehensive basis for model evaluation and future benchmarking.
- This work systematically evaluates a diverse set of machine learning models, including SVR, GBR, XGBoost, RFR, KNN, LR, and ANN on complex geotechnical problem.
- The study demonstrates the superior performance of ensemble learning techniques, in accurately predicting pollution diffusion.

2. RELATED WORKS

Numerous studies utilized advancements in numerical methods and machine learning (Abbas J. Al-Taie and Al-Bayati 2021), and optimization techniques (Delpisheh et al. 2024) for accurate prediction and modeling of pollutant diffusion in soils. It is worth mentioning that traditional approaches such as physical and empirical models struggled with nonlinear interactions in complex environmental systems (Akesh 2017). On the other hand, machine learning techniques have proven their effectiveness in addressing these challenges by capturing nonlinear and multimodal relationships in environmental datasets (Zhu, Yang, and Ren 2023), they developed a 3D coupled soil-groundwater model for heavy metal transport and demonstrating its efficiency in predicting contamination and optimizing treatment at sites (Zhang et al. 2024). A study (Samborska-Goik and Pogrzeba 2024) critically reviewed modeling tools for organic contaminant emphasizing the importance of accurate tools for reactive transport simulations. Another study reviewed the applications of the ML in groundwater quality modeling and highlighted ensemble methods compared to traditional

techniques (Karimi et al. 2025). ML approaches (Wu et al. 2024) was implemented to study pollutant motion in aquifer systems further underscoring their practical applicability in environmental engineering. On optimization algorithms (Chen 2024), a study had developed machine learning models to predict soil phthalate pollution and highlighted key environmental factors for effective risk assessment (Pan et al. 2024).

These studies highlighted the growing convergence of ML and optimization in environmental sciences.

3. SOIL CONTAMINATION DATASET

The dataset used in this study describes the diffusion of contaminant materials on soil surface and can be downloaded at:

https://www.kaggle.com/datasets/shaymaaalsamia/diffusion-contamination

The dataset consists of five columns:

1. C which is the concentration; a measure of contamination concentration.
2. K is the permeability which is the permeability of the soil in terms of contamination flow.
3. k ratio is a permeability ratio that might represent the relative comparison of different materials or layers.
4. t represents the duration of observation or contamination progression.
5. h is the output, representing the vertical penetration of contamination in the soil.
6. D is another output representing horizontal diffusion of contamination in the soil.

Here, we have inputs that span various units (concentration, permeability, ratios, and time) and outputs in terms of both vertical and horizontal contamination spread where the multimodal nature of the dataset arises from the variety of variables. The values of "C" have a wide range impacting vertical and horizontal diffusion through its interaction with permeability. The "K" and "k-ratios" suggest varying conditions of soil permeability, introducing multiple dimensions to predict the behavior of contamination spread. Also, "t" evolves non-linearly, which likely adds time-dependent complexity to both outputs "h" and "D".

Fig. 1 shows that concentration shows a slight positive correlation with both h (0.14) and D (0.18). This implies that as the concentration of contaminants increases, there is a little tendency for vertical depth and horizontal diffusion to increase, but the relationship is not strong. Permeability has a high positive correlation with h (0.87), indicating that as the permeability of the soil increases, the vertical depth of contamination penetration rises significantly. K has a medium correlation with D (0.60), suggesting that permeability also affects horizontal diffusion but not as strongly as vertical penetration. k ratio shows a medium positive correlation with h

(0.58) with a weak correlation to D (0.30). This means the permeability ratio contributes more to the vertical depth of contamination than the horizontal spread. Time has a moderate positive correlation with h (0.62), implying that the vertical depth of penetration increases over time. The t has a weaker but still positive correlation with D (0.45), indicating that the horizontal diffusion also increases over time but to a lesser extent than vertical depth. There is a moderate positive correlation between h and D (0.52), meaning that as contamination penetrates deeper into the soil vertically, it also tends to spread more horizontally. This suggests an interdependent relationship between the two forms of contamination spread.

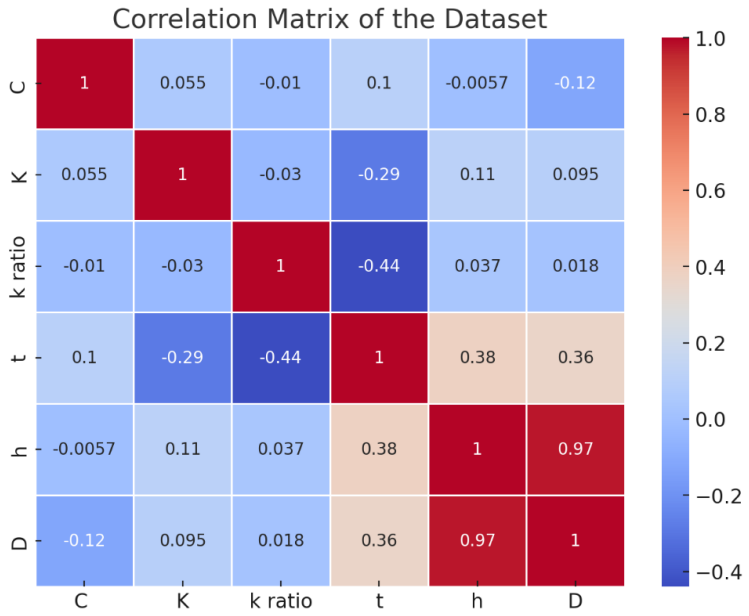

Fig. 1 Correlation matrix of the soil contamination dataset

Table 1 shows the average value of each variable (mean). For example, the average concentration C is 201,544.12, while the average vertical depth h is 1.81. Std (Standard Deviation) indicates the degree to which the data deviates from the mean. A larger standard deviation refers to that the values are more spread out. For example, the standard deviation for C is very large (166,173.38), indicating a wide range of concentration values in the dataset. Min (Minimum) is the smallest value observed for each variable. For example, the minimum vertical depth h is 0, and the minimum horizontal diffusion D is 2. 25% (1st Quartile) is the 25th percentile, also called the first quartile. It means that 25% of the data falls under this value. For instance, 25% of the vertical depth h observations are below 0.98, and 25% of D observations are below 2.83. 50% (Median) is the 50th percentile (also called the median), where half of the data is above this value and half is below. It provides the central tendency of the dataset. For example, the median vertical depth h is 1.58, and the median horizontal diffusion D is 3.35. 75% (3rd Quartile) is the 75th percentile, meaning 75% of the data falls under this value. For

instance, 75% of the vertical depth h observations are below 2.43, and 75% of the D observations are below 3.85. Max is the largest value observed for each variable. For example, the maximum vertical depth h is 4.28, and the maximum horizontal diffusion D is 4.85.

Table 1 Statistical Summary of the Soil Contamination Dataset, Including Key Variables and Their Descriptive Statistics

	C	K	k ratio	t	h	D
count	680	680	680	680	680	680
mean	201544.1176	0.000137379	0.609558824	182.2675245	1.808957353	3.449745588
Standard deviation	166173.378	9.00E-05	0.446369527	375.5897111	0.918146877	0.667362821
Minimum	25000	3.70E-05	0.1	0	0	2
25%	100000	3.70E-05	0.1	10	0.98	2.83
50%	100000	0.000107	1	30.41666667	1.965	3.51
75%	450000	0.000236	1	169.0416667	2.6	4.04
Maximum	450000	0.000248	1	3333.333333	3.51	4.66

4. PREDICTIVE MODELS

4.1. Support Vector Regression

Support vector regression is an adaptation of support vector machines and designed to handle regression tasks. It utilizes part from the training dataset, known as support vectors, to construct a regression hyperplane within defined threshold called the epsilon tube. The SVR is highly effective for small to medium-sized datasets, it seeks to minimize the margin of error and model complexity to avoid overfitting. The kernel trick in SVR enables it to capture nonlinear relationships, making it suitable for complex datasets where linear models maybe failed. Fig. 2 shows the actual diffusion of the contamination against the predicted diffusion using the SVR method.

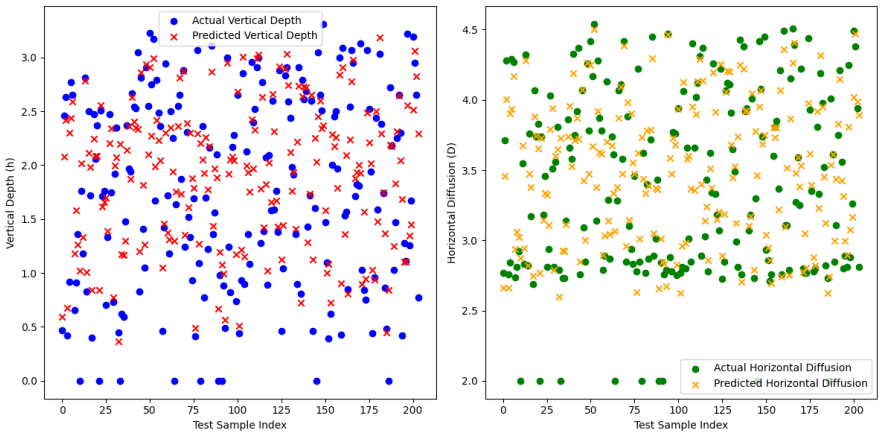

Fig. 2 Predicted vs. actual pollution diffusion using SVR

4.2. Random Forest Regression

Random forest regression is an ensemble learning technique, builds multiple decision trees during the training phase and combines their outputs to generate predictions for regression tasks. The RFR is an approach that addresses the overfitting issues seen in the individual

decision trees by taking the average of the outputs of several trees. Each tree is trained on a unique subset of the dataset, and the final output is obtained by taking the average of all predictions. Random forest is well-regarded for its reliability and capacity to manage high-dimensional data and intricate relationships among input features. Fig. 3 displays the actual diffusion of the contamination alongside the predicted diffusion produced by the RFR model.

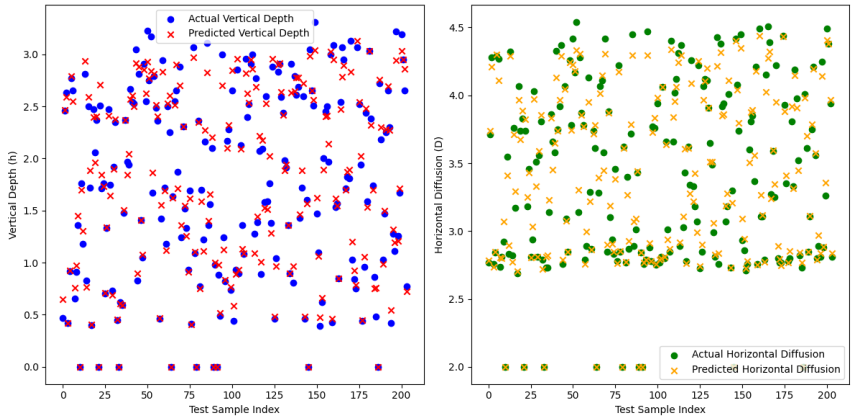

Fig. 3 Predicted vs. Actual pollution diffusion using random forest regression

4.3. Gradient Boosting Regression

Gradient boosting regression is an ensemble method that sequentially constructs regression trees with each tree addressing the mistakes made by its predecessor. GBR transforms weak learners into a robust predictive model by repeatedly refining the predictions based on the gradient of the error function. The GBR is highly effective for capturing complex nonlinear relationships and provides more accurate forecasts than many other methods. Fig. 4 illustrates the actual diffusion of the contamination against the predicted using the GBR method.

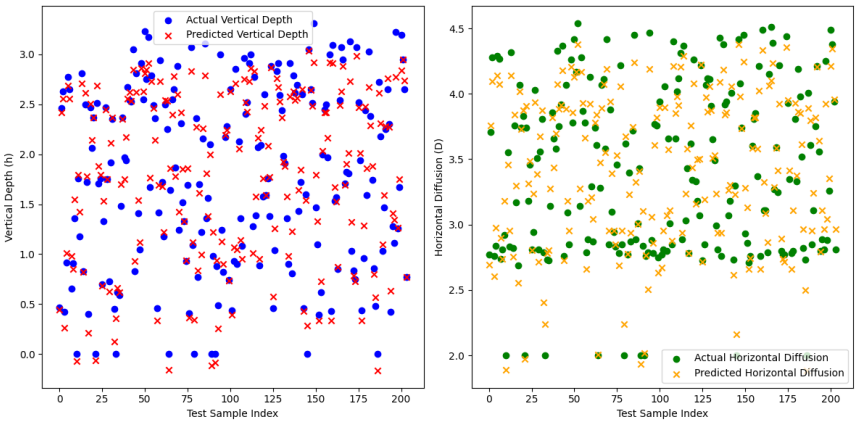

Fig. 4 Predicted vs. Actual Pollution Diffusion Using GBR

4.4. XGBoost Regression

XGBoost is an advanced variant of gradient boosting, it is fine-tuned for enhanced speed and efficiency and integrate regularization methods to minimze overfitting and incorporating better mechanisms for managing missing data. Fig.5 presents the actual diffusion of the contamination vs. the predicted diffusion by employing the XGBoost regression.

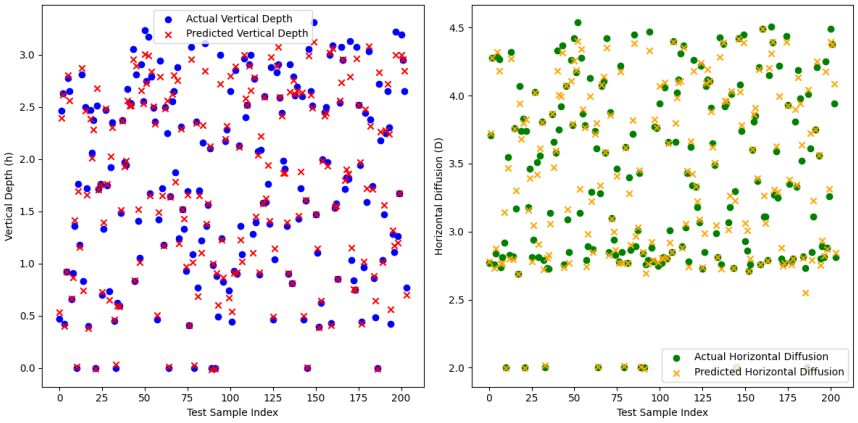

Fig. 5 Predicted vs. actual pollution diffusion using XGBoost regression

4.5. K-Nearest Neighbors

K-nearest neighbors regression is a non-parametric technique where the prediction for a data point is determined by averaging the outputs of its k-nearest neighbors. Fig.6 compares the actual contamination diffusion with the predicted diffusion obtained using KNN regression.

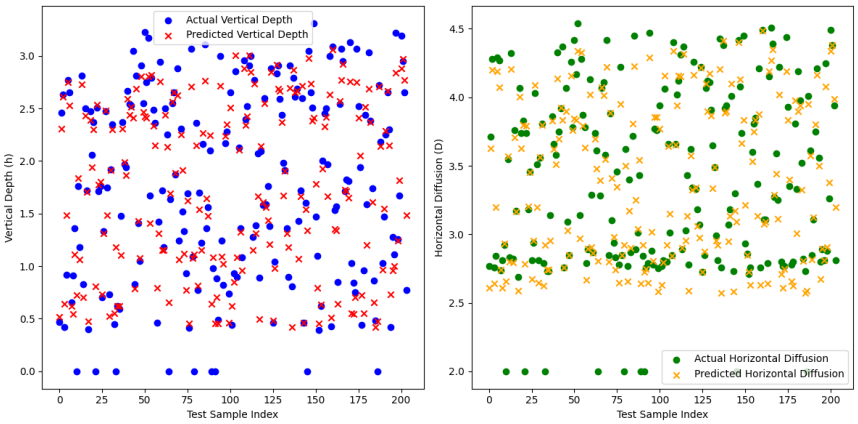

Fig.6 Predicted vs. actual pollution diffusion using KNN

4.6. Linear Regression

Linear regression is among the simplest and most highly interpretable models used in predictive modeling. However, its primary limitation is that it struggles with complex nonlinear relationships and multimodal data Fig. 7 reveals the actual diffusion of the contamination against the predicted one using the LR method.

4.7. Artificial Neural Networks ANN

The ANN is versatile and can model patterns in data by introducing non-linearity through activation functions. ANN require large amounts of data and computational resources and they are prone to overfitting without careful regularization. Fig.8 presents the resulting prediction of the diffusion using the ANN method.

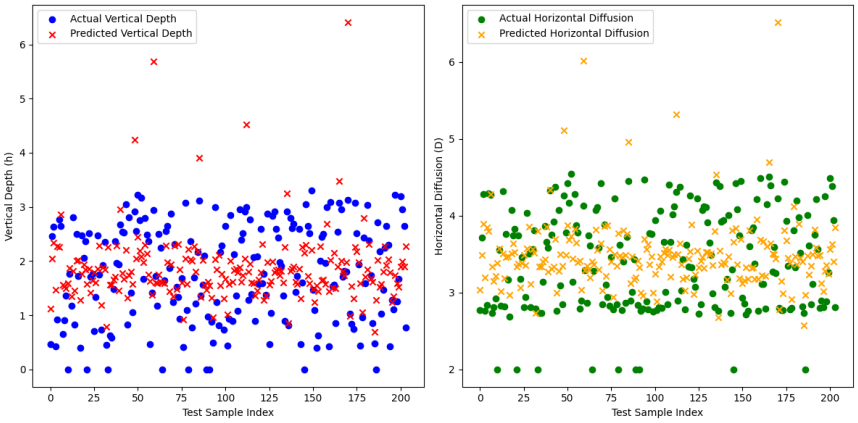

Fig. 7 Predicted vs. Actual pollution diffusion using linear regression

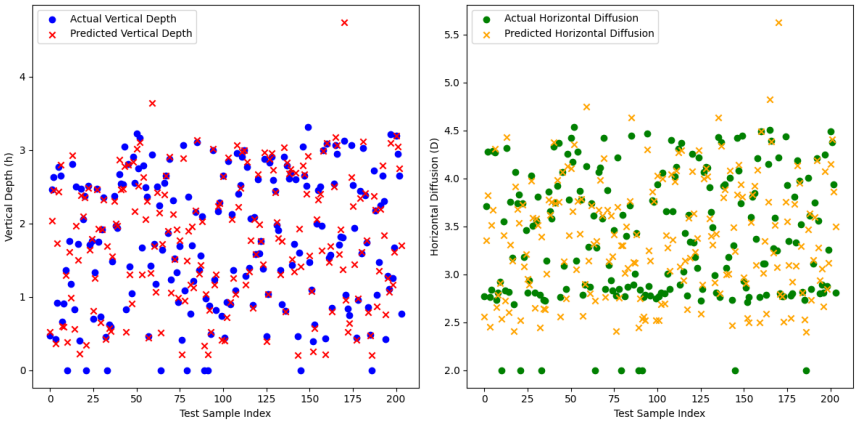

Fig. 8 Predicted vs. actual pollution diffusion using ANN

5. EVALUATION METRICS

To evaluate the performance of the machine learning models used in this study, we utilized the following metrics that provide a comprehensive assessment of the accuracy and reliability of the predictive models.

1. MSE that measures the average squared difference between the predicted and observed values; it penalizes larger errors more significantly and expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (1)$$

where n is the number of data points, y_i is the actual value, $\hat{y}_i$ is the predicted value.

2. RMSE which is the square root of MSE, and calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

3. MAE which calculates the average absolute difference between the predicted and observed values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

4. The R^2metric, which is also known as the coefficient of determination, measures the proportion of the variance in the observed values that is predictable from the independent variables. This metric ranges from 0 to 1, with higher values indicating better model performance and described as follows

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

where $\bar{y}$ is the mean of actual values.

6. RESULTS AND DISCUSSIONS

In this section, we present and analyze the outcomes of various machine learning algorithms employed to estimate pollution's vertical depth and horizontal spread. Figs. 3 through 9 illustrate the comparison between the observed and forecasted values of pollution diffusion across different models, while Tables 2 and 3 provide a summary of the evaluation criteria for both depth and width predictions, including the following metrics: MSE, RMSE, MAE, and R-squared (R^2). SVR performed moderately well, capturing the overall trend of the pollution diffusion data but with noticeable deviations, especially in the more extreme cases. The MSE for vertical depth prediction was 0.26051 with an R^2 of 0.68252, indicating that while the model explains around 68% of the variance, there is significant room for improvement. Similarly, the MSE for horizontal diffusion was 0.13798 with an R^2 of 0.67646, suggesting similar performance for both dimensions.The Random Forest Regression model demonstrated outstanding predictive performance, as reflected in Fig. 3. The predictions closely match the actual data points, with minimal scatter between actual and predicted values. This is confirmed by the evaluation metrics in Table 2, where the MSE for depth prediction was 0.01226, and the R^2 was 0.98504, suggesting that the model captured 98.5% of the variance in the data. For horizontal diffusion, the performance was similarly high, with an MSE of 0.00704 and an R^2 of 0.98348. This makes RFR one of the best-performing models for both dimensions, effectively reducing errors and providing highly accurate predictions.

GBR, as seen in Fig.4, also produced highly accurate predictions but slightly more deviations than Random Forest. The model performed well, At an MSE of 0.02991 and an R^2 of 0.96354 for depth prediction and an MSE of 0.02528 and an R^2 of 0.94070 for diffusion prediction. This indicates that GBR can effectively capture the nonlinearities in the dataset, but it is slightly less robust than Random Forest in terms of generalization. However, it remains one of the top-performing models.

The XGBoost Regression model Fig. 5 delivered one of the best performances in this study. The model's predictive capability is reflected by its extremely low MSE values, particularly for

the depth prediction, where the MSE was 0.00885, and the R^2 was 0.98921. The MSE was 0.00540 for horizontal diffusion, and the R^2 was 0.98733. These results indicate that XGBoost is highly effective in handling the complexity and multimodal nature of the dataset, offering excellent predictive performance in both dimensions. The scatter plots in Fig.5 show a very tight clustering of the actual and predicted values, confirming its accuracy.

The performance of KNN regression, shown in Fig. 6, was comparatively lower than that of tree-based models like Random Forest and XGBoost. The MSE for vertical depth was 0.07785, with an R^2 of 0.90512, while the horizontal diffusion predictions yielded an MSE of 0.05069 and an R^2 of 0.88114. Although KNN does not perform as well as gradient boosting methods, it still captures significant patterns in the data, albeit with more errors, particularly at extreme values. This can be observed in the scattering of predicted points around the actual values in Fig. 6.

LR was the poorest-performing model, as expected for a dataset with high non-linearity and multimodality. Fig.7 shows that the model fails to capture the variance in the data, particularly at extreme values, with large deviations between the actual and predicted points. The MSE for depth prediction was 0.67342, with a very low R^2 of 0.17933, indicating that LR explains only 17.9% of the variance in the data. Similarly, the horizontal diffusion predictions were poor, with an MSE of 0.35934 and an R^2 of 0.15747. This highlights the limitations of linear models in capturing complex relationships present in this dataset.

The ANN, as shown in Fig. 8, performed moderately well but still left room for improvement. The predictions show a noticeable scatter around the actual values, and while the model captures the general trend, it struggles with some extreme values. The MSE for vertical depth prediction was 0.11908 with an R^2 of 0.85488, while for horizontal diffusion, the MSE was 0.11132 with an R^2 of 0.73897. Although the model outperforms LR and KNN, it falls short compared to tree-based ensemble models like XGBoost and Random Forest.

From the analysis of Figs. 2 through 8 and Tables 2 and 3, it is evident that ensemble methods, particularly Random Forest, Gradient Boosting, and XGBoost, provide superior performance in predicting both vertical depth and horizontal diffusion of pollution. XGBoost achieved the best overall results, with the lowest MSE and highest R^2 in both dimensions, closely followed by Random Forest and Gradient Boosting.

In brief, ensemble learning models like XGBoost and random forest are recommended for predicting pollution diffusion in complex and multimodal datasets due to their robustness, high accuracy, and ability to generalize well to unseen data.

Table 2 Performance evaluation of machine learning models for predicting vertical depth of soil contamination

Method	MSE	RMSE	MAE	R^2
SVR	0.26051	0.51040	0.33948	0.68252
RFR	0.01226	0.11076	0.08018	0.98504
GBR	0.02991	0.17296	0.13473	0.96354
XGBoost	0.00885	0.09409	0.06384	0.98921
KNN	0.07785	0.27902	0.18174	0.90512
Linear Regression	0.67342	0.82062	0.65507	0.17933
ANN	0.11908	0.34508	0.18152	0.85488

Table 3 Performance evaluation of machine learning models for predicting horizontal diffusion of soil contamination

Method	MSE	RMSE	MAE	R^2
SVR	0.13798	0.37146	0.23620	0.67646
RFR	0.00704	0.08393	0.06088	0.98348
GBR	0.02528	0.15902	0.12789	0.94070
XGBoost	0.00540	0.07350	0.05214	0.98733
KNN	0.05069	0.22515	0.13756	0.88114
Linear Regression	0.35934	0.59945	0.47615	0.15747
ANN	0.11132	0.33366	0.21284	0.73897

Models like ANN and KNN require further optimization or specific conditions to match the accuracy of ensemble methods. The LR, is fundamentally unsuitable for the nonlinear complexity of this study.

7. CONCLUSION

This study evaluates the efficiency of different machine learning models in predicting both the vertical penetration and horizontal spread of soil contamination. By employing a multimodal dataset, the work demonstrated that ensemble methods; XGBoost and random forest provide the most accurate and reliable predictions. Their ability to capture nonlinear and multimodal relationships highlights their value as practical tools for environmental modeling. Gradient boosting also showed strong performance, though with slightly higher sensitivity to parameter tuning.

The results revealed the limitations of linear regression, which failed to account for the dataset inherent complexity. The models of KNN and ANN their performance suggests that further optimization would be needed to bring them closer to ensemble-level accuracy. Overall, the findings emphasize the importance of advanced learning techniques in tackling the challenges of soil pollution prediction.

8. REFERENCES

Abbas J. Al-Taie, Abbas J Al-Taie, and Ahmed Al-Bayati. 2021. “APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO PREDICT SOIL RECOMPRESSION INDEX

AND RECOMPRESSION RATIO." Kufa Journal of Engineering 9 (4 SE-Peer-reviewed Articles): 246–57. https://doi.org/10.30572/2018/KJE/090417.

Akesh, Ammar Ashour. 2017. "Analytical Study for Heavy Metals Pollution in Surface Water and Sediment for Selected Rivers of Basrah Governorate." Kufa Journal of Engineering 8 (2): 105–18.

Al-Ani, Mohammad M, Mohammad Y Fattah, and Mahmoud T A Al-Lamy. 2009. "Artificial Neural Networks Analysis of Treatment Process of Gypseous Soils." Engineering and Technology Journal 27 (9): 1811–32.

Albedran, Hazim, and Károly Jármai. 2023. "Evolutionary Control System of Asymmetric Quadcopter." International Review of Applied Sciences and Engineering 14 (3): 374–82. https://doi.org/https://doi.org/10.1556/1848.2022.00584.

Albedran, Hazim, Shaymaa Alsamia, and Edina Koch. 2025. "Flower Fertilization Optimization Algorithm with Application to Adaptive Controllers." Scientific Reports 15 (1): 6273. https://doi.org/10.1038/s41598-025-89840-1.

Albedran, Hazim. 2025. "Advanced Model Predictive Control Optimization for Automotive Dynamics." In International Conference on Intelligent and Fuzzy Systems, 3–11. Springer.

Alsamia, S. and Koch, E., 2024. Random forest regression on pullout resistance of a pile. Pollack Periodica, 19(3), pp.28-33.

Alsamia, Shaymaa M, Mohammed S Mahmood, and Ali Akhtarpour. 2020. "Prediction of the Contamination Track in Al-Najaf City Soil Using Numerical Modelling." In IOP Conference Series: Materials Science and Engineering, 888:12050. IOP Publishing.

Alsamia, Shaymaa, and Edina Koch. 2023. "EVALUATION THE BEHAVIOR OF PULLOUT FORCE AND DISPLACEMENT FOR A SINGLE PILE: EXPERIMENTAL VALIDATION WITH PLAXIS 3D." Kufa Journal of Engineering 14 (2): 105–16.

Alsamia, Shaymaa, Edina Koch, and Hanaa Shihab Hamadi. 2023. "Comparative Study of Metaheuristics on Optimal Design of Gravity Retaining Wall." Pollack Periodica. https://doi.org/https://doi.org/10.1556/606.2023.00826.

Alsamia, Shaymaa, Hazim Albedran, and Károly Jármai. 2022. "Comparative Study of Different Metaheuristics on CEC 2020 Benchmarks." In Vehicle and Automotive Engineering 4: Select Proceedings of the 4th VAE2022, Miskolc, Hungary, 709–19. Springer. https://doi.org/https://doi.org/10.1007/978-3-031-15211-5_59.

Alsamia, Shaymaa, Hazim Albedran, and Mohammed Sh Mahmood. 2022. “Contamination Depth Prediction in Sandy Soils Using Fuzzy Rule-Based Expert System.” International Review of Applied Sciences and Engineering.

Amiri, Mohammad, Masoud Dehghani, Tohid Javadzadeh, and Sepideh Taheri. 2022. “Effects of Lead Contaminants on Engineering Properties of Iranian Marl Soil from the Microstructural Perspective.” Minerals Engineering 176: 107310.

Bosu, Subrajit, Natarajan Rajamohan, Su Shiung Lam, and Yasser Vasseghian. 2023. “Environmental Remediation of Agrochemicals and Dyes Using Clay Nanocomposites: Review on Operating Conditions, Performance Evaluation, and Machine Learning Applications.” Reviews of Environmental Contamination and Toxicology 261 (1): 17.

Chen, I-Chun. 2024. “Predicting Regional Sustainable Development to Enhance Decision-Making in Brownfield Redevelopment Using Machine Learning Algorithms.” Ecological Indicators 163: 112117.

Delpisheh, Mostafa, Benyamin Ebrahimpour, Abolfazl Fattahi, Majid Siavashi, Hamed Mir, Hossein Mashhadimoslem, Mohammad Ali Abdol, Mina Ghorbani, Javad Shokri, and Daniel Niblett. 2024. “Leveraging Machine Learning in Porous Media.” Journal of Materials Chemistry A.

Ghafil, H.N., and K. Jármai. 2019. “Optimum Dynamic Analysis of a Robot Arm Using Flower Pollination Algorithm.” In Advances and Trends in Engineering Sciences and Technologies III-Proceedings of the 3rd International Conference on Engineering Sciences and Technologies, ESaT 2018. https://doi.org/https://doi.org/10.1201/9780429021596.

Ghafil, Hazim Nasir, and Károly Jármai. 2020. “Optimization Algorithms for Inverse Kinematics of Robots with MATLAB Source Code.” In Vehicle and Automotive Engineering, 468–77. Springer.

Ghafil, Hazim Nasir, Kovács László, and Károly Jármai. 2019. “Investigating Three Learning Algorithms of a Neural Networks during Inverse Kinematics of Robots.” Solutions for Sustainable Development, 33–40. https://doi.org/https://doi.org/10.1201/9780367824037.

Haggerty, Ryan, Jianxin Sun, Hongfeng Yu, and Yusong Li. 2023. “Application of Machine Learning in Groundwater Quality Modeling-A Comprehensive Review.” Water Research 233: 119745.

Jalghaf, Humam Kareem, Ali Habeeb Askar, Hazim Albedran, Endre Kovács, and Károly Jármai. 2023. "Comparative Study of Different Meta-Heuristics on Optimal Design of a Heat Exchanger." Pollack Periodica 18 (2): 119–24.

Karimi, Hadi, Soheil Sahour, Matin Khanbeyki, Vahid Gholami, Hossein Sahour, Sina Shahabi-Ghahfarokhi, and Mohsen Mohammadi. 2025. "Enhancing Groundwater Quality Prediction through Ensemble Machine Learning Techniques." Environmental Monitoring and Assessment 197 (1): 1–25.

Li, Xiaonuo, Shiyi Yi, Andrew B Cundy, and Weiping Chen. 2022. "Sustainable Decision-Making for Contaminated Site Risk Management: A Decision Tree Model Using Machine Learning Algorithms." Journal of Cleaner Production 371: 133612.

Liu, Xian, Dawei Lu, Aiqian Zhang, Qian Liu, and Guibin Jiang. 2022. "Data-Driven Machine Learning in Environmental Pollution: Gains and Problems." Environmental Science & Technology 56 (4): 2124–33.

Luo, Jiannan, Xi Ma, Yefei Ji, Xueli Li, Zhuo Song, and Wenxi Lu. 2023. "Review of Machine Learning-Based Surrogate Models of Groundwater Contaminant Modeling." Environmental Research, 117268.

Maghawry, Ahmed, Rania Hodhod, Yasser Omar, and Mohamed Kholief. 2021. "An Approach for Optimizing Multi-Objective Problems Using Hybrid Genetic Algorithms." Soft Computing 25: 389–405.

Martins, Joaquim R R A, and Andrew Ning. 2021. Engineering Design Optimization. Cambridge University Press.

Mayer, Martin János, Artúr Szilágyi, and Gyula Gróf. 2020. "Environmental and Economic Multi-Objective Optimization of a Household Level Hybrid Renewable Energy System by Genetic Algorithm." Applied Energy 269: 115058.

Obead, Imad Habeeb, Hassan Ali Omran, and Mohammed Yousif Fattah. 2021. "Implementation of Artificial Neural Network to Predict the Permeability and Solubility Models of Gypseous Soil." Pertanika Journal of Science & Technology 29 (1).

Pan, Boyou, Jialin Lei, Bogui Pan, Hong Tian, and Li Huang. 2024. "Dialogue between Algorithms and Soil: Machine Learning Unravels the Mystery of Phthalates Pollution in Soil." Journal of Hazardous Materials, 136604.

Salim, Abdulrahman A, Zainab B Mohammed, and Mohammed Y Fattah. 2022. “Influence of Adding Plant Fly Ash on the Geotechnical Properties and Pollution of Sanitary Landfill Soil.” Engineering and Technology Journal 40 (11): 1385–98.

Samborska-Goik, Katarzyna, and Marta Pogrzeba. 2024. “A Critical Review of the Modelling Tools for the Reactive Transport of Organic Contaminants.” Applied Sciences 14 (9): 3675.

Tahmasebi, Pejman, Serveh Kamrava, Tao Bai, and Muhammad Sahimi. 2020. “Machine Learning in Geo-and Environmental Sciences: From Small to Large Scale.” Advances in Water Resources 142: 103619.

Wu, Yingdong, Jiang Yu, Zhi Huang, Yinying Jiang, Zixin Zeng, Lei Han, Siwei Deng, and Jie Yu. 2024. “Migration of Total Petroleum Hydrocarbon and Heavy Metal Contaminants in the Soil–Groundwater Interface of a Petrochemical Site Using Machine Learning: Impacts of Convection and Diffusion.” RSC Advances 14 (44): 32304–13.

Xiang, Song, Xiaosong He, Qi Yang, and Yuxin Wang. 2024. “Migration and Natural Attenuation of Leachate Pollutants in Bedrock Fissure Aquifer at a Valley Landfill Site.” Environmental Pollution, 124963.

Xue, Shengguo, Wenshun Ke, Jiaqing Zeng, Carlito Baltazar Tabelin, Yi Xie, Lu Tang, Chao Xiang, and Jun Jiang. 2023. “Pollution Prediction for Heavy Metals in Soil-Groundwater Systems at Smelting Sites.” Chemical Engineering Journal 473: 145499.

Ye, Zhiping, Jiaqian Yang, Na Zhong, Xin Tu, Jining Jia, and Jiade Wang. 2020. “Tackling Environmental Challenges in Pollution Controls Using Artificial Intelligence: A Review.” Science of the Total Environment 699: 134279.

Zhang, Hai-li, Peng Zhao, Wen-yan Gao, Bao-hua Xiao, Xue-feng Yang, Lei Song, Xiang Feng, Lin Guo, Yong-ping Lu, and Hai-feng Li. 2024. “Contaminant Transport Modelling of Heavy Metal Pollutants in Soil and Groundwater: An Example at a Non-Ferrous Smelter Site.” Journal of Central South University 31 (4): 1092–1106.

Zhu, Jun-Jie, Meiqi Yang, and Zhiyong Jason Ren. 2023. “Machine Learning in Environmental Research: Common Pitfalls and Best Practices.” Environmental Science & Technology 57 (46): 17671–89.