frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 10 May 2023
DOI 10.3389/fmicb.2023.1152059

OPEN ACCESS

EDITED BY
Muhammad Shaaban,
Bahauddin Zakariya University, Pakistan

REVIEWED BY
Alex Furman,
Technion Israel Institute of Technology, Israel
Murray Close,
Institute of Environmental Science and Research (ESR), New Zealand

*CORRESPONDENCE
Xijuan Chen
✉ chenxj@iae.ac.cn

RECEIVED 27 January 2023
ACCEPTED 25 April 2023
PUBLISHED 10 May 2023

CITATION
Chen F, Zhou B, Yang L, Chen X and Zhuang J (2023) Predicting bacterial transport through saturated porous media using an automated machine learning model. *Front. Microbiol.* 14:1152059. doi: 10.3389/fmicb.2023.1152059

Predicting bacterial transport through saturated porous media using an automated machine learning model

Fengxian Chen[1], Bin Zhou[2], Liqiong Yang[1], Xijuan Chen[1]* and Jie Zhuang[3]

[1]Key Laboratory of Pollution Ecology and Environmental Engineering, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, Liaoning, China, [2]Faculty of Medicine, University of Augsburg, Augsburg, Germany, [3]Department of Biosystems Engineering and Soil Science, Center for Environmental Biotechnology, The University of Tennessee, Knoxville, TN, United States

Escherichia coli, as an indicator of fecal contamination, can move from manure-amended soil to groundwater under rainfall or irrigation events. Predicting its vertical transport in the subsurface is essential for the development of engineering solutions to reduce the risk of microbiological contamination. In this study, we collected 377 datasets from 61 published papers addressing *E. coli* transport through saturated porous media and trained six types of machine learning algorithms to predict bacterial transport. Eight variables, including bacterial concentration, porous medium type, median grain size, ionic strength, pore water velocity, column length, saturated hydraulic conductivity, and organic matter content were used as input variables while the first-order attachment coefficient and spatial removal rate were set as target variables. The eight input variables have low correlations with the target variables, namely, they cannot predict target variables independently. However, using the predictive models, input variables can effectively predict the target variables. For scenarios with higher bacterial retention, such as smaller median grain size, the predictive models showed better performance. Among six types of machine learning algorithms, Gradient Boosting Machine and Extreme Gradient Boosting outperformed other algorithms. In most predictive models, pore water velocity, ionic strength, median grain size, and column length showed higher importance than other input variables. This study provided a valuable tool to evaluate the transport risk of *E.coli* in the subsurface under saturated water flow conditions. It also proved the feasibility of data-driven methods that could be used for predicting other contaminants' transport in the environment.

KEYWORDS

bacterial transport, automated machine learning, first-order attachment coefficient, spatial removal rate, machine learning

Highlights

- The predictive models showed better performance when bacterial retention was high.
- Spatial removal rate is a better target variable than first-order attachment coefficient.
- Algorithms based on gradient boosting outperformed other machine learning algorithms.

1. Introduction

Microbiologically contaminated drinking water is estimated to cause 485,000 diarrhoeal deaths each year (WHO, 2022). Manure-borne pathogens, as an important source of microbiological contamination, can be transported from surface soil to groundwater through manure disposal, storage, and land application, posing a threat to public health (Pachepsky et al., 2006; Alegbeleye and Santana, 2020). The vertical transport of bacteria in soil is an important route for microbiological contamination, especially in the area with shallow groundwater level (Guo et al., 2020). Numerous flow-through experiments, from lab-scale to field-scale, were conducted to identify the key parameters and explore the mechanism of bacterial transport through porous media or natural soils (Schinner et al., 2010; Bradford et al., 2013; Yang et al., 2019; Oudega et al., 2021; He et al., 2022). The results showed that the bacterial transport process was controlled by environmental factors (e.g., soil texture, particle size, soil surface charges, organic matter content, water content, ionic strength, water flow velocity) and bacterial properties (e.g., concentration, cell size, cell surface charge, hydrophobicity) (Sepehrnia et al., 2017; Zhong et al., 2017). Based on these experimental data, mathematical models, such as attachment/detachment model and filtration theory, were used to describe and compare the bacterial transport behaviors (Šimunek et al., 2012). These studies are rigorous enough to emphasize some specific factors of bacterial transport in each independent study. However, they have two obvious limitations. One is the difficulty to evaluate many variables at one time, specifically, as the number of variables increases, the workload of experiments will grow exponentially. Second, the model parameters from one specific experiment may be not suitable for predicting bacterial transport in other scenarios because of different soil properties, scales, and water flow conditions. Therefore, data-driven methods beyond the experiments are needed for the prediction of bacterial transport under a wide range of scenarios.

In recent years, data-driven methods have been developed fast due to the advances in machine learning (Solomatine et al., 2009; Hiemer and Zapperi, 2021). Machine learning is the process of generating models that learning from historic data to make predictions for the future or other scenarios (Samanpour et al., 2018). Previous bacterial flow-through experiments produced large amount of data, providing a foundation for developing data-driven models such as machine learning. However, the machine learning approach demands a high level of technical sophistication in model selection and hyperparameter tuning, constituting an obstacle for non-machine learning experts (Hutter et al., 2019). Thus, automated machine learning (AML) was proposed in recent years. The AML approach can automatically optimize machine learning process by integrating feature engineering, model selection, hyperparameter optimization, and model evaluation (Waring et al., 2020). The recent AML models include AutoWEKA (Kotthoff et al., 2019), Auto-sklearn (Feurer et al., 2020), AutoGluon (Erickson et al., 2020), H2O AutoML (LeDell and Poirier, 2020), and TPOT (Olson and Moore, 2016). They were used for analysis and prediction in many areas of environmental science, such as groundwater redox conditions, landslide hazard analysis, methane production in anaerobic digestion, and groundwater radioactivity (Wilson et al., 2020; Qi et al., 2021; Fallatah et al., 2022; Xu et al., 2022). However, they have not been applied to predict environmental fates of pathogens and pollutants.

The first step for AML modeling is to define the input variables (known as features) and target variables (Waring et al., 2020). For the vertical transport of bacteria in soil, the input variables can be bacterial properties and environmental factors, while the target variables should be parameters that can effectively describe the transport and retention behaviors of bacteria using the advection-dispersion equation as follows,

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} - kC \tag{1}$$

In the equation, C (cell mL^{-1}) is the bacterial concentration, t (h) is time, x (cm) is travel distance, D is dispersion coefficient (cm^2 h^{-1}), v (cm h^{-1}) is the pore water velocity and k (h^{-1}) is first-order attachment coefficient (Kretzschmar et al., 1997). For a given scenario, the parameter C, t, x, D, v can be easily measured or calculated. If k can be predicted as a target variable, the bacterial transport and retention can be described. The simplified bacterial retention in soil can be written as

$$\frac{dC}{dt} = -kC \tag{2}$$

This exponential decay function describes that the bacterial concentration decreases at a rate proportional to its current concentration value, and k is an exponential decay constant. Besides, the bacterial removal can also be described from distance perspective,

$$\frac{dC}{dx} = -\lambda C \tag{3}$$

where λ (cm^{-1}) is spatial removal rate of bacteria (Kretzschmar et al., 1997; Pang, 2009). The exponential decay function describes that the bacterial concentration decreases along the distance at a rate proportional to its current concentration value, and λ is an exponential decay constant.

Both parameters k and λ describe the bacterial transport and retention in soil. The former is time based, while the latter is distance based. The relation between k and λ under steady state flow is $k = \lambda v$.

Escherichia coli (*E. coli*) is commonly used as a common indicator of fecal contamination (Odonkor and Ampofo, 2013). In this study, we extracted 377 datasets of *E. coli* vertical transport in saturated porous media from 61 papers. The objective of this study was to predict the two main parameters of *E. coli* transport in saturated porous media using an AML model. The input variables were bacterial concentration, porous medium type, median grain size, ionic strength, pore water velocity, column length, saturated hydraulic conductivity, and organic matter content, and the target variables were first-order attachment coefficient and spatial removal rate. Based on the AML model (H2O AutoML), six types of machine learning algorithms and 20 predictive models were trained, and their performances were evaluated.

2. Materials and methods

2.1. Data collection and analysis

A literature search was conducted on the Web of Science to collect data regarding the *E. coli* transport in soil/sand (Supplementary Table S1). Reducing the number of input variables is beneficial for improving the model prediction performance. From the collected literature, we selected eight most frequently investigated factors as input variables, including bacterial concentration, porous medium type, median grain size, ionic strength, pore water velocity, column length, saturated hydraulic conductivity, and organic matter content. Some bacterial properties, such as cell size (with a length of 1–2 μm and a radius of 0.5 μm) and hydrophobicity (mostly hydrophilic), were not considered because their difference among different *E. coli* strains was small. The bacterial zeta potential (−10 ~ −50 mV) was not considered because it was strongly correlated with the ionic strength of liquid phase. Among the eight input variables, porous medium type was categorical variable (i.e., sand, intact soil, and disturbed soil), and the other seven variables were numerical variables.

The target variable (the first-order attachment coefficient k) was collected through the following ways: (1) when the first-order attachment model (Equation 1) was used to fit break-through curves, k was an optimized parameter; (2) if the breakthrough curves were not fitted with models or not fitted by the first-order attachment model, k was converted from λ based on $k = \lambda v$.

The target variable (spatial removal rate λ) was collected through the following two ways. First, in breakthrough curves, when the effluent concentration reached a plateau, λ was calculated from the following equation:

$$\lambda = -\frac{1}{L}\ln\left(\frac{C_f}{C_0}\right) \tag{4}$$

where L is the length of the soil column, C_0 is the bacterial input concentration, and C_f is the effluent concentration at the plateau of the breakthrough curve (Kretzschmar et al., 1997). Second, for the breakthrough curves without a plateau, λ was calculated using the following equation:

$$\lambda = -\frac{1}{L}\ln\left(\frac{M_{\text{eff}}}{M_{\text{in}}}\right) \tag{5}$$

where is M_{in} the total injected bacterial mass and M_{eff} is the total effluent bacterial mass (Kretzschmar et al., 1997).

If k is easily obtained, λ is converted by $k = \lambda v$. In the study, the values of k and λ was extracted from the literature or otherwise calculated from the breakthrough curves.

2.2. Automated machine learning model

We used an automated machine learning model, H2O AutoML (LeDell and Poirier, 2020). The H2O AutoML integrated many common machine learning algorithms, including Deep Learning, Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Extremely Randomized Trees (XRT), and Extreme Gradient Boosting (XGBoost) (LeDell and Poirier, 2020). The H2O AutoML provides some function calls for automatically training the candidate models. The codes in R for modeling training and model performance evaluation were shown in the Supplementary material. The variables and machine learning model training process are shown in Figure 1.

2.3. Model performance measures

The collected bacterial transport datasets were randomly divided into two datasets: 80% training dataset and 20% test dataset. The training dataset was used to train the machine learning models and the test dataset was used to evaluate the model performance. 5-fold cross-validation was performed. Based on the test dataset and the model predicted target variables, three statistical measures were used to evaluate the model performance: root mean squared error (RMSE), mean absolute error (MAE) and absolute relative residual (ARR). For each predicted value, there is one ARR value (Li et al., 2020).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{6}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{7}$$

$$\text{ARR} = \left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{8}$$

where $\hat{y}_i$ is the model predicted value, y_i is the observed value, and n is the number of data points in the test dataset.

2.4. Explainable analysis

In H2O AutoML, 20 machine learning algorithms was synchronously examined, and their ranking was listed on a leaderboard based on their performance. Variable importance and Shapley additive explanations (SHAP) were used to analyze the importance and contribution of the input variables (LeDell and Poirier, 2020). The variable importance is ranged from 0 to 100% and represents the importance of each input variable for the target variable. SHAP shows the contribution of each variable in each row of data and the trend of the variable's influence, i.e., positive, or negative influence (Lundberg and Lee, 2017).

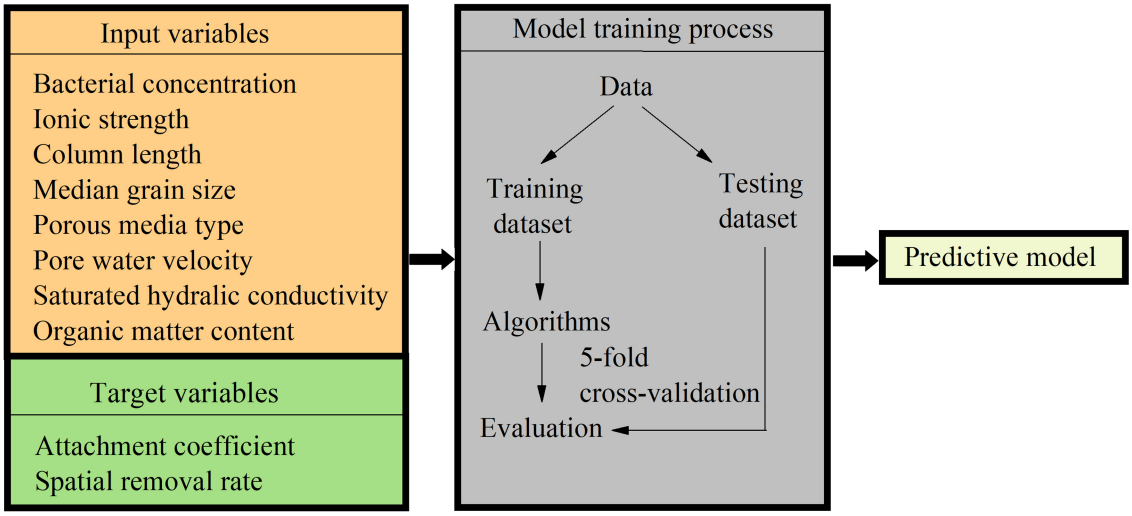

FIGURE 1
The variables and machine learning model training process in this study.

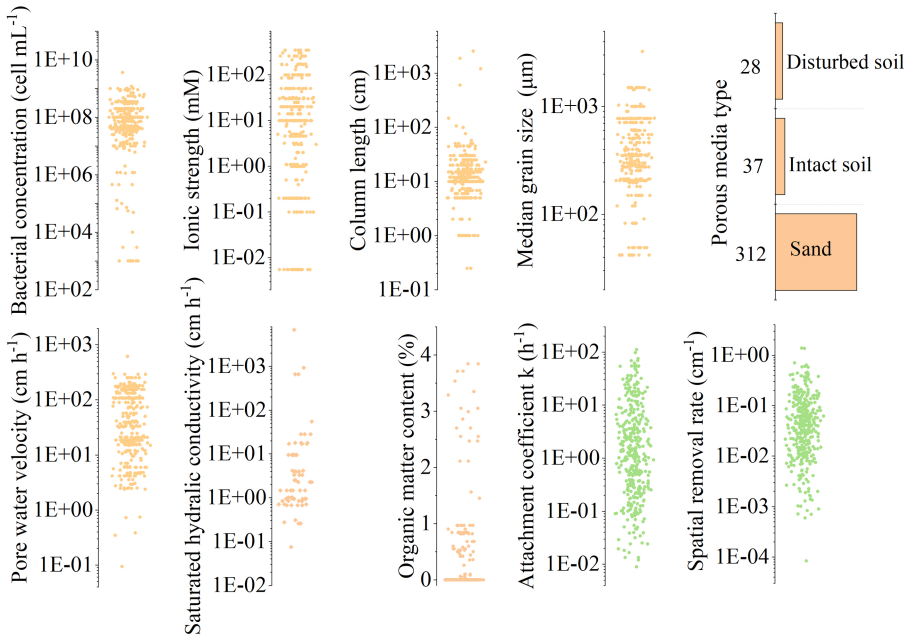

FIGURE 2
Statistics of model input and target variables datasets.

3. Results and discussion

3.1. Overview of the collected datasets

The column scatter plot of collected dataset is shown in Figure 2. For the input variables, the bacterial concentration ranged from 10^3 to 10^9 cell mL^{-1}; the ionic strength of liquid phase was normalized as NaCl solution ranging from 10^{-3} to 10^3 mM; the column length was ranged from 0.25 to 2,565 cm; the median grain sizes of porous media varied from 42 to 1,500 μm; the pore water velocity ranged from 0.09 to 618 cm h^{-1}; the saturated hydraulic conductivity ranged from 0.08 to 55.8 cm h^{-1}; the organic matter content ranged from 0.05 to 3.84%; the number of datasets for sand, intact soil, and disturbed soil was 314, 37, and 28, respectively. Because literature usually described particle size distribution of sand and soil in different ways, i.e., median grain size is usually used for sand; while soil texture is described by sand, silt, and clay percentage (here we use saturated hydraulic conductivity to reflect soil texture property). Thus, in our datasets, most sand lacked saturated hydraulic conductivity data and most soil lack median grain sizes data. Regarding the target variables, the first-order attachment coefficient (k) ranged from 0.009 to 111.6 h^{-1}, and the spatial removal rate (λ) ranged from 0.00008 to 1.379 cm^{-1}. Overall,

the collected datasets covered the range in a variety of flow-through experiments.

The Pearson correlation coefficient (r) between pairwise variables is shown in Figure 3. The r values of k and λ with other eight input variables ranged from −0.25 to 0.34 and − 0.22 to 0.17, respectively. The weak Pearson correlations imply that the eight input variables could not be a valid predictor independently. The r between k and λ was 0.68, indicating that these two target variables were moderately correlated in positive manner ($k = \lambda v$). However, when v varies, the increased k does not necessarily indicate an increase in λ.

Determination of variables and data collection are the two key prerequisites for training machine learning models. In the study, two types of factors were not considered. The first type of factors has small differences but may influence bacterial transport behaviors. For example, cell surface characteristics (e.g., physiological state, flagella type, and extracellular polymeric substances) were reported to affect bacterial transport in sand or soil (Madumathi et al., 2017; Du et al., 2021; Zhang et al., 2021). These variables were not included into the AML model because they were not clearly defined in most literature. The second type may affect bacterial transport, but the effects are minor, such as temperature and bacterial starvation (Kim and Walker, 2009; Walczak et al., 2012).

3.2. Model performance

Based on the results of 5-fold cross-validation, the correlation between predicted values and observed values for the test datasets was plotted in Figures 4A,B. During the H2O AutoML training process, 20 models (six types of machine learning algorithms) were trained simultaneously, and their ranking based on RMSE and MAE was listed in Supplementary Table S2. For the predictions of k and λ, the best model was GBM, followed by XGboost. The slope of linear fitting in GBM was close to 1. The R^2 of linear regression using GBM was 0.82 and 0.85 in Figures 4A,B, respectively. XRT, Deep learning and DRF showed similar performance, belonging to the second tier. GLM showed the worst performance among six machine learning algorithms.

The data points were plotted in log scale, some negative values in predicted values cannot be shown in figures. The summary of negative values was as follows. The number of negative values in predicting k with GBM, XGboost, XRT, Deep learning, DRF and GLM was 14, 31, 0, 9, 0, and 17, respectively. The number of negative values in predicting λ with GBM, XGboost, XRT, Deep learning, DRF and GLM was 4, 19, 0, 4, 0, and 13, respectively. The numbers of negative values indicated that although XGboost showed similar performance as GBM in the statistical measures (RSME, MAE, R^2), many negative predicted values greatly affected its reliability. Compared to the prediction of k, prediction of λ showed fewer negative values.

Although GBM showed the best model performance, the hyperparameters affected the model performance greatly. In Supplementary Table S2, the number that attached to each machine learning algorithm refers to different hyperparameter settings, which can be extracted from the H2O AutoML for further tuning in a regular machine learning model. For example, the GBM_5 was ranked first while the GBM_1 was ranked eighteenth for predicting k. The RMSE of GBM_1 was two times of that of GBM_5. Compared with regular machine learning model training, the automatically set hyperparameters in the H2O AutoML significantly improved the efficiency of model training (LeDell and Poirier, 2020).

To better analyze the results from predictive models, we chose the best performed model GBM to continue the statistical analysis. The absolute relative residuals (ARR) in GBM predictive model for predicting k and λ are shown in Figures 5A,B. Because the target

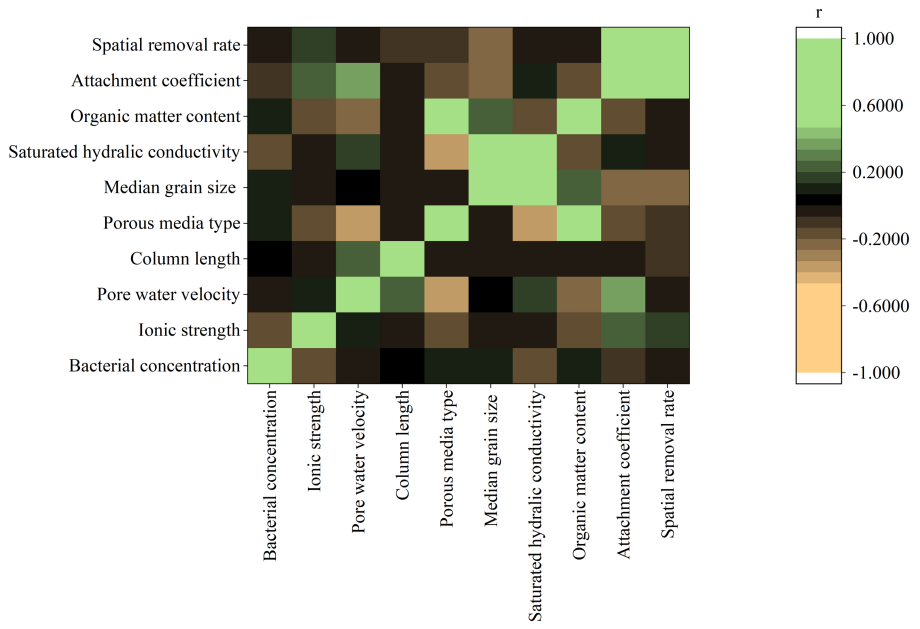

FIGURE 3
Pearson product-moment correlation coefficient (r) between pairwise variables (porous media type: 1-sand; 2-distubed soil; 3-intact soil).

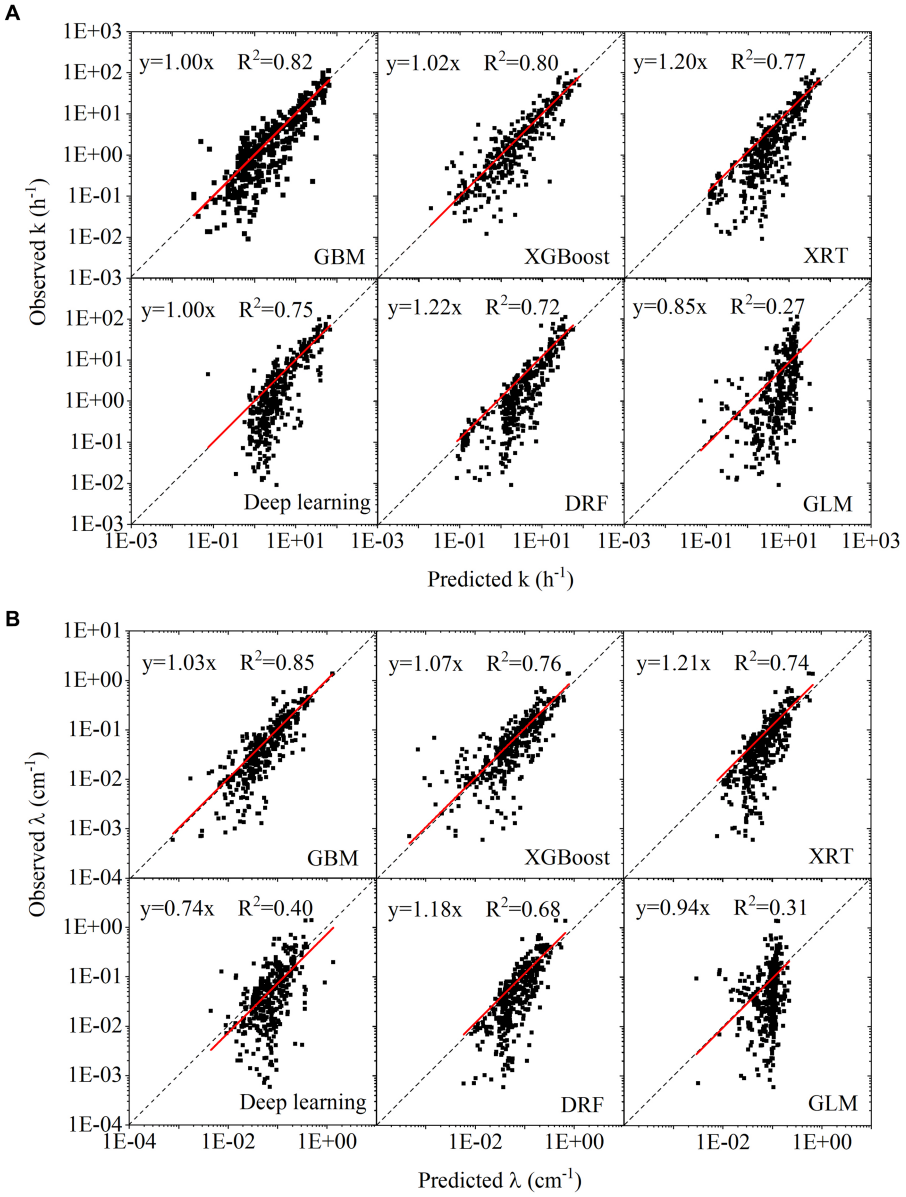

FIGURE 4
(A) Linear regression between observed target variables and predicted target variables for first-order attachment coefficient (*k*); **(B)** linear regression between observed target variables and predicted target variables for spatial removal rate (*λ*).

variable values were widely distributed (cover 4–5 orders of magnitude), the basic statistical measures (RMSE, MAE, R^2) cannot reflect the errors of all the values equally. In contrast, the ARR value can provide more information (Li et al., 2020). As the increase of observed values, the ARR decreased. Namely, when k and λ was high, the ARR was relatively small. Specifically, for scenarios with higher bacterial retention, such as smaller median grain size, higher ionic strength, and longer bacterial transport distance, the predictive models showed better performance. Compared to the ARR values for predicting k, that for predicting λ were slightly lower. It indicates that predicting λ is a better choice than predicting k. Besides, in Figures 5A,B, the ARR values of three types of porous media were separately plotted. It showed that there was no obvious difference among sand, intact soil, and disturbed soil.

3.3. Analysis of variable importance

The importance of variables in six types of machine learning algorithms is shown in Figures 6A,B. The variable importance for predicting k and λ was similar, ranking as pore water velocity = ionic strength > median grain size > column length > bacterial concentration > organic content > porous medium type = saturated hydraulic conductivity.

The SHAP contribution of GBM is demonstrated in Figures 7A,B. In the SHAP summary plot, the color represents the normalized values of each data point in the testing dataset. The SHAP contribution value represents the positive or negative contribution of each data point for predicting target variables. As shown in the SHAP summary plot, larger median grain size and column length contributed

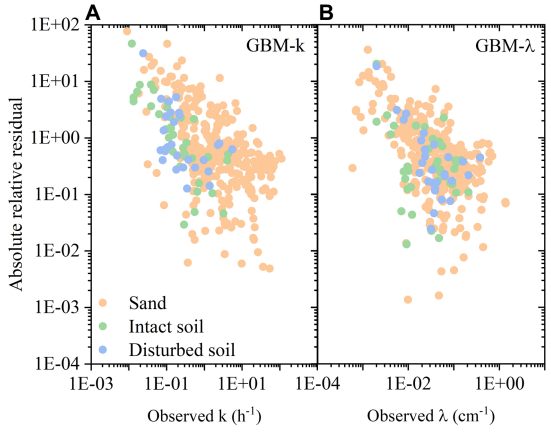

FIGURE 5
(A) Absolute relative residual in GBM predictive model for predicting first-order attachment coefficient (***k***); **(B)** Absolute relative residual in GBM predictive model for predicting spatial removal rate (***λ***).

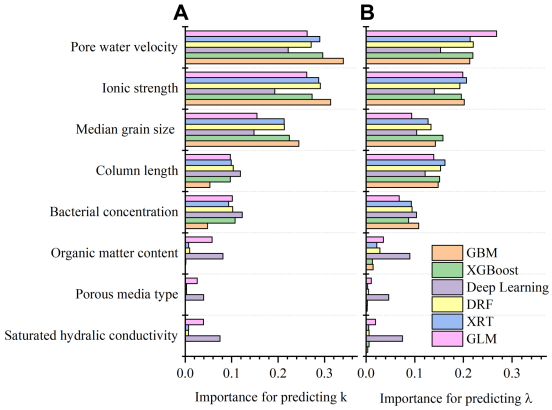

FIGURE 6
(A) Variable importance in six predictive models for predicting first-order attachment coefficient (***k***); **(B)** variable importance in six predictive models for predicting spatial removal rate (***λ***).

to smaller k and λ value, larger ionic strength led to bigger k and λ value, and bacterial concentration, organic matter content, porous medium type and saturated hydraulic conductivity did not influence k and λ value. Besides, pore water velocity showed opposite effect on k and λ value because of the inverse relation (i.e., $k = \lambda v$) between k and λ when pore water velocity is a constant.

Pore water velocity showed highest importance in the predictions of k and λ. This is because bacterial transport could be increased by increasing pore water velocity (Bradford et al., 2006; Choi et al., 2007; Chen et al., 2022). Higher pore water velocity is accompanied by higher water shear force and less bacteria-soil contact time, which can reduce bacterial mechanical filtration and bacterial attachment, respectively (Hendry et al., 1999; Li et al., 2005; Syngouna and Chrysikopoulos, 2011). The responses of k and λ to pore water velocity agree with the filtration theory (Logan et al., 1995). Ionic strength is another important variable. Higher ionic strength favored bacterial retention as manifested by larger k and λ values. The responses of k and λ to solution ionic strength agree with the DLVO theory (Kim and Walker, 2009; Walczak et al., 2012; Wang et al., 2013; Bai et al., 2017).

Median grain size of porous media is correlated to soil texture and soil porosity. The bigger median grain size of porous media was favorable to reduce bacterial retention (i.e., smaller k and λ). This trend has been well confirmed in previous studies (Gannon et al., 1991; Balkhair, 2017; Sepehrnia et al., 2017). The effect of soil column length can be regarded as a scale effect (Hijnen et al., 2005; Knappett et al., 2014). As shown in the SHAP contribution figures, the upscaling of bacterial transport resulted in smaller k and λ. The distribution of column length data points was more scattered for predicting λ than predicting k, suggesting that λ is more sensitive to column length. This result is consistent with a recent study that the upscaling effect is more pronounced for λ than k (Oudega et al., 2021).

For porous medium type, bacterial concentration, organic content, and saturated hydraulic conductivity, the SHAP contribution shows relatively aggregated distribution. Many studies showed that the intact soil could greatly facility bacterial transport because of preferential flow in macropore-dominated pathways (McLeod et al., 1998; Safadoust et al., 2011; Chen et al., 2022). Nevertheless, the effect of intact soil or disturbed soil was not obvious in our predictive models. The reason may be the database of soil was small, which is not enough for distinguishing the contribution of intact or disturbed soil. Similarly, the contribution of organic content, and saturated hydraulic conductivity faced with the same problem (small database limited the variable importance analysis). Previous studies showed that increase in bacterial concentration may either increase or decrease bacterial retention by blocking or ripening, respectively (Bradford and Bettahar, 2006; Zhang et al., 2010). This concentration effect may also be related to solution ionic strength (Bradford et al., 2009). Therefore, it is not possible to conclude the positive or negative contribution of bacterial concentration to the transport.

3.4. Comparison of different machine learning algorithms

Among six machine learning algorithms, GBM and XGBoost belongs to gradient boosting algorithms; DRF and XRT are random forest-based algorithms; Deep Learning is based on artificial neural networks; GLM is a flexible generalization of ordinary linear regression (Mahesh, 2020). From the perspective of regression performance, the algorithms based on gradient boosting outperformed other algorithms, because they can optimize on different loss functions to make the function fit very flexible. Thus, they have higher accuracy than random forest-based algorithms, artificial neural networks and generalized linear model (Madeh Piryonesi and El-Diraby, 2021). The GLM always showed the worst performance, implying that GLM has weak ability to predict when the problem is complicated (many variables).

Although GBM and XGBoost showed better performance than other algorithms, some predicted values from them were negative, which violates the physics of bacterial transport and retention in porous media. In contrast, the random forest-based algorithms, such as DRF and XRT, are more consistent with physical laws. Therefore, to predict bacterial transport parameters, different machine learning algorithms may be used in a combination way. For example, under most conditions, GBM and XGBoost are suitable. Once the negative values appear, the random forest-based algorithms, such as DRF and XRT can be used as a supplement.

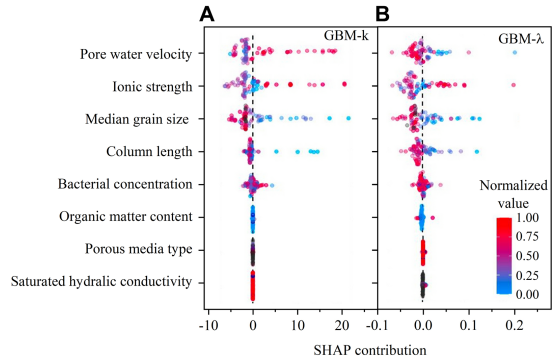

FIGURE 7
(A) SHAP contribution in GBM predictive model for predicting first-order attachment coefficient (k); **(B)** SHAP contribution in GBM predictive model for predicting spatial removal rate (λ).

3.5. Limitations and future application

The main limitation of the predictive models is that the datasets of soil in database is small (17.2%), which makes that predictive models were not sensitive to porous medium type, saturated hydraulic conductivity, and organic matter content. This study only collected data from the transport of *E coli*. Therefore, the data in literature was not enough. Further studies, such as building a bigger database comprising of different bacteria or microorganisms, may provide more extensive and accurate predictive models.

Compared to previous studies based on controlled variables, the data-driven machine learning algorithms provides an advantageous approach for regression problems. With machine learning algorithms, many input variables that have low correlations with the target variables can predict the target variables with very high accuracy. This extraordinary performance of the AML model has been confirmed in other studies (Wilson et al., 2020; Qi et al., 2021; Fallatah et al., 2022; Xu et al., 2022).

The collected datasets and R code for the H2O AutoML are shown in the Supplementary material. The users may use the collected datasets and their own data to train an AML model and then use it to predict transport parameters for *E. coli* under saturated flow conditions. When combined with mechanism-based models and software, such as Hydrus, the bacterial transport can be simulated and visualized (Šimunek et al., 2012). The users may also add more variables to expand to a more comprehensive prediction for solute and colloid transport in the vadose zone.

4. Conclusion

In this study, literature-based data regarding *E. coli* transport through saturated sand or soil were used to train an AML model (H2O AutoML). We used bacterial concentration, porous medium type, median grain size, ionic strength, pore water velocity, column length, saturated hydraulic conductivity, and organic matter content as input variables to predict first-order attachment coefficient (k) and spatial removal rate (λ). The results showed that the trained machine learning models were reliable tools to predict key parameters for *E. coli* transport through saturated porous media. Among six types of machine learning algorithms, the gradient boosted based algorithms, such as Gradient Boosting Machine and Extreme Gradient Boosting, outperformed other machine learning algorithms. The predictive models showed better performance when bacterial retention was high. Besides, spatial removal rate is a better target variable than first-order attachment coefficient. Compared with traditional controlled variable experiments, the data-driven AML accomplished the goal that predicting bacterial transport from a comprehensive perspective. This approach offers a new way of thinking for predicting environmental fates of various pollutants.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JZ and XC contributed to conception and design of the study. FC organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. BZ, LY, JZ, and XC reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by the National Natural Science Foundation of China (Grant No. 41730858), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA28090100), and the Liaoning Science and Technology Plan Project (Grant No. 2021JH2/10300079).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1152059/full#supplementary-material

References

Alegbeleye, O. O., and Santana, A. S. (2020). Manure-borne pathogens as an important source of water contamination: an update on the dynamics of pathogen survival/transport as well as practical risk mitigation strategies. *Int. J. Hyg. Environ. Health* 227:113524. doi: 10.1016/j.ijheh.2020.113524

Bai, H., Cochet, N., Pauss, A., and Lamy, E. (2017). DLVO, hydrophobic, capillary and hydrodynamic forces acting on bacteria at solid-air-water interfaces: their relative impact on bacteria deposition mechanisms in unsaturated porous media. *Colloids Surf. B: Biointerfaces* 150, 41–49. doi: 10.1016/j.colsurfb.2016.11.004

Balkhair, K. S. (2017). Modeling fecal bacteria transport and retention in agricultural and urban soils under saturated and unsaturated flow conditions. *Water Res.* 110, 313–320. doi: 10.1016/j.watres.2016.12.023

Bradford, S. A., and Bettahar, M. (2006). Concentration dependent transport of colloids in saturated porous media. *J. Contam. Hydrol.* 82, 99–117. doi: 10.1016/j.jconhyd.2005.09.006

Bradford, S. A., Kim, H. N., Haznedaroglu, B. Z., Torkzaban, S., and Walker, S. L. (2009). Coupled factors influencing concentration-dependent colloid transport and retention in saturated porous media. *Environ. Sci. Technol.* 43, 6996–7002. doi: 10.1021/es900840d

Bradford, S. A., Morales, V. L., Zhang, W., Harvey, R. W., Packman, A. I., Mohanram, A., et al. (2013). Transport and fate of microbial pathogens in agricultural settings. *Crit. Rev. Environ. Sci. Technol.* 43, 775–893. doi: 10.1080/10643389.2012.710449

Bradford, S. A., Simunek, J., Bettahar, M. E. H. D. I., Van Genuchten, M. T., and Yates, S. R. (2006). Significance of straining in colloid deposition: evidence and implications. *Water Resour. Res.* 42:W12S15. doi: 10.1029/2005WR004791

Chen, J., Yang, L., Chen, X., Ripp, S., and Zhuang, J. (2022). Coupled effects of pore water velocity and soil heterogeneity on bacterial transport: intact vs. repacked soils. *Front. Microbiol.* 13:730075. doi: 10.3389/fmicb.2022.730075

Choi, N. C., Kim, D. J., and Kim, S. B. (2007). Quantification of bacterial mass recovery as a function of pore-water velocity and ionic strength. *Res. Microbiol.* 158, 70–78. doi: 10.1016/j.resmic.2006.09.007

Du, M., Wang, L., Ebrahimi, A., Chen, G., Shu, S., Zhu, K., et al. (2021). Extracellular polymeric substances induced cell-surface interactions facilitate bacteria transport in saturated porous media. *Ecotoxicol. Environ. Saf.* 218:112291. doi: 10.1016/j.ecoenv.2021.112291

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., et al. (2020). Autogluon-tabular: robust and accurate automl for structured data. *arXiv preprint arXiv: 2003.06505*

Fallatah, O., Ahmed, M., Gyawali, B., and Alhawsawi, A. (2022). Factors controlling groundwater radioactivity in arid environments: an automated machine learning approach. *Sci. Total Environ.* 830:154707. doi: 10.1016/j.scitotenv.2022.154707

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F. (2020). Auto-sklearn 2.0: the next generation. *Arxiv preprint arXiv:2007.04074, 24*

Gannon, J. T., Mingelgrin, U., Alexander, M., and Wagenet, R. J. (1991). Bacterial transport through homogeneous soil. *Soil Biol. Biochem.* 23, 1155–1160. doi: 10.1016/0038-0717(91)90028-I

Guo, X., Hu, H., Meng, H., Liu, L., Xu, X., and Zhao, T. (2020). Vertical distribution and affecting factors of *Escherichia coli* over a 0–400 cm soil profile irrigated with sewage effluents in northern China. *Ecotoxicol. Environ. Saf.* 205:111357. doi: 10.1016/j.ecoenv.2020.111357

He, L., Li, M., Wu, D., Guo, J., Zhang, M., and Tong, M. (2022). Freeze-thaw cycles induce diverse Bacteria release behaviors from quartz sand columns with different water saturations. *Water Res.* 221:118683. doi: 10.1016/j.watres.2022.118683

Hendry, M. J., Lawrence, J. R., and Maloszewski, P. (1999). Effects of velocity on the transport of two bacteria through saturated sand. *Groundwater* 37, 103–112. doi: 10.1111/j.1745-6584.1999.tb00963.x

Hiemer, S., and Zapperi, S. (2021). From mechanism-based to data-driven approaches in materials science. *Mater. Theory* 5, 1–9. doi: 10.1186/s41313-021-00027-3

Hijnen, W. A., Brouwer-Hanzens, A. J., Charles, K. J., and Medema, G. J. (2005). Transport of MS2 phage, *Escherichia coli*, *Clostridium perfringens*, Cryptosporidium parvum, and Giardia intestinalis in a gravel and a sandy soil. *Environ. Sci. Technol.* 39, 7860–7868. doi: 10.1021/es050427b

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Cham, Switzerland: Springer Nature, 219.

Kim, H. N., and Walker, S. L. (2009). *Escherichia coli* transport in porous media: influence of cell strain, solution chemistry, and temperature. *Colloids Surf. B: Biointerfaces* 71, 160–167. doi: 10.1016/j.colsurfb.2009.02.002

Knappett, P. S. K., Du, J., Liu, P., Horvath, V., Mailloux, B. J., Feighery, J., et al. (2014). Importance of reversible attachment in predicting *E. coli* transport in saturated aquifers from column experiments. *Adv. Water Resour.* 63, 120–130. doi: 10.1016/j.advwatres.2013.11.005

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K. (2019). "Auto-WEKA: automatic model selection and hyperparameter optimization in WEKA" in *Automated machine learning* (Cham: Springer), 81–95.

Kretzschmar, R., Barmettler, K., Grolimund, D., Yan, Y. D., Borkovec, M., and Sticher, H. (1997). Experimental determination of colloid deposition rates and collision efficiencies in natural porous media. *Water Resour. Res.* 33, 1129–1137. doi: 10.1029/97WR00298

LeDell, E., and Poirier, S. (2020). "H2o automl: scalable automatic machine learning" in *Proceedings of the auto ML workshop at ICML*, vol. 2020, Vienna, Austria.

Li, W., Meng, P., Hong, Y., and Cui, X. (2020). Using deep learning to preserve data confidentiality. *Appl. Intell.* 50, 341–353. doi: 10.1007/s10489-019-01515-3

Li, X., Zhang, P., Lin, C. L., and Johnson, W. P. (2005). Role of hydrodynamic drag on microsphere deposition and re-entrainment in porous media under unfavorable conditions. *Environ. Sci. Technol.* 39, 4012–4020. doi: 10.1021/es048814t

Logan, B. E., Jewett, D. G., Arnold, R. G., Bouwer, E. J., and O'Melia, C. R. (1995). Clarification of clean-bed filtration models. *J. Environ. Eng.* 121, 869–873. doi: 10.1061/(ASCE)0733-9372(1995)121:12(869)

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Proces. Syst.* 30, 4768–4777. doi: 10.48550/arXiv.1705.07874

Madeh Piryonesi, S., and El-Diraby, T. E. (2021). Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. *J. Infrastruct. Syst.* 27:04021005. doi: 10.1061/(ASCE)IS.1943-555X.0000602

Madumathi, G., Philip, L., and Bhallamudi, S. M. (2017). Transport of *E. coli* in saturated and unsaturated porous media: effect of physiological state and substrate availability. *Sādhanā* 42, 1007–1024. doi: 10.1007/s12046-017-0650-8

Mahesh, B. (2020). Machine learning algorithms-a review. *Int. J. Sci. Res. (IJSR) [Internet]* 9, 381–386. doi: 10.21275/ART20203995

McLeod, M., Schipper, L. A., and Taylor, M. D. (1998). Preferential flow in a well drained and a poorly drained soil under different overhead irrigation regimes. *Soil Use Manag.* 14, 96–100. doi: 10.1111/j.1475-2743.1998.tb00622.x

Odonkor, S. T., and Ampofo, J. K. (2013). *Escherichia coli* as an indicator of bacteriological quality of water: an overview. *Microbiol. Res.* 4:e2. doi: 10.4081/mr.2013.e2

Olson, R. S., and Moore, J. H. (2016). "TPOT: a tree-based pipeline optimization tool for automating machine learning" in *Workshop on automatic machine learning* (New York, USA: PMLR), 66–74.

Oudega, T. J., Lindner, G., Derx, J., Farnleitner, A. H., Sommer, R., Blaschke, A. P., et al. (2021). Upscaling transport of *Bacillus subtilis* endospores and Coliphage phi X174 in heterogeneous porous media from the column to the field scale. *Environ. Sci. Technol.* 55, 11060–11069. doi: 10.1021/acs.est.1c01892

Pachepsky, Y. A., Sadeghi, A. M., Bradford, S. A., Shelton, D. R., Guber, A. K., and Dao, T. (2006). Transport and fate of manure-borne pathogens: modeling perspective. *Agric. Water Manag.* 86, 81–92. doi: 10.1016/j.agwat.2006.06.010

Pang, L. (2009). Microbial removal rates in subsurface media estimated from published studies of field experiments and large intact soil cores. *J. Environ. Qual.* 38, 1531–1559. doi: 10.2134/jeq2008.0379

Qi, W., Xu, C., and Xu, X. (2021). AutoGluon: a revolutionary framework for landslide hazard analysis. *Nat. Hazards Res.* 1, 103–108. doi: 10.1016/j.nhres.2021.07.002

Safadoust, A., Mahboubi, A. A., Gharabaghi, B., Mosaddeghi, M. R., Voroney, P., Unc, A., et al. (2011). Bacterial filtration rates in repacked and weathered soil columns. *Geoderma* 167-168, 204–213. doi: 10.1016/j.geoderma.2011.08.014

Samanpour, A. R., Ruegenberg, A., and Ahlers, R. (2018). "The future of machine learning and predictive analytics" in *Digital marketplaces unleashed* (Berlin, Heidelberg: Springer), 297–309.

Schinner, T., Letzner, A., Liedtke, S., Castro, F. D., Eydelnant, I. A., and Tufenkji, N. (2010). Transport of selected bacterial pathogens in agricultural soil and quartz sand. *Water Res.* 44, 1182–1192. doi: 10.1016/j.watres.2008.11.038

Sepehrnia, N., Memarianfard, L., Moosavi, A. A., Bachmann, J., Guggenberger, G., and Rezanezhad, F. (2017). Bacterial mobilization and transport through manure enriched soils: experiment and modeling. *J. Environ. Manag.* 201, 388–396. doi: 10.1016/j.jenvman.2017.07.009

Šimunek, J., Van Genuchten, M. T., and Šejna, M. (2012). HYDRUS: Model use, calibration, and validation. *Trans. ASABE* 55, 1263–1276. doi: 10.13031/2013.42239

Solomatine, D., See, L. M., and Abrahart, R. J. (2009). Data-driven modelling: concepts, approaches and experiences. *Pract. Hydroinformat.* 68, 17–30. doi: 10.1007/978-3-540-79881-1_2

Syngouna, V. I., and Chrysikopoulos, C. V. (2011). Transport of biocolloids in water saturated columns packed with sand: effect of grain size and pore water velocity. *J. Contam. Hydrol.* 126, 301–314. doi: 10.1016/j.jconhyd.2011.09.007

Walczak, J. J., Wang, L., Bardy, S. L., Feriancikova, L., Li, J., and Xu, S. (2012). The effects of starvation on the transport of *Escherichia coli* in saturated porous media are dependent on pH and ionic strength. *Colloids Surf. B: Biointerfaces* 90, 129–136. doi: 10.1016/j.colsurfb.2011.10.010

Wang, Y., Bradford, S. A., and Šimůnek, J. (2013). Transport and fate of microorganisms in soils with preferential flow under different solution chemistry conditions. *Water Resour. Res.* 49, 2424–2436. doi: 10.1002/wrcr.20174

Waring, J., Lindvall, C., and Umeton, R. (2020). Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* 104:101822. doi: 10.1016/j.artmed.2020.101822

Wilson, S. R., Close, M. E., Abraham, P., Sarris, T. S., Banasiak, L., Stenger, R., et al. (2020). Achieving unbiased predictions of national-scale groundwater redox conditions via data oversampling and statistical learning. *Sci. Total Environ.* 705:135877. doi: 10.1016/j.scitotenv.2019.135877

World Health Organization (WHO). (2022). *Drinking-water (fact sheet)*. Available at: https://www.who.int/news-room/fact-sheets/detail/drinking-water

Xu, R. Z., Cao, J. S., Ye, T., Wang, S. N., Luo, J. Y., Ni, B. J., et al. (2022). Automated machine learning-based prediction of microplastics induced impacts on methane production in anaerobic digestion. *Water Res.* 223:118975. doi: 10.1016/j.watres.2022.118975

Yang, L., Chen, X., Zeng, X., Radosevich, M., Ripp, S., Zhuang, J., et al. (2019). Surface-adsorbed contaminants mediate the importance of chemotaxis and haptotaxis for bacterial transport through soils. *Front. Microbiol.* 10:2691. doi: 10.3389/fmicb.2019.02691

Zhang, M., He, L., Jin, X., Bai, F., Tong, M., and Ni, J. (2021). Flagella and their properties affect the transport and deposition behaviors of *escherichia coli* in quartz sand. *Environ. Sci. Technol.* 55, 4964–4973. doi: 10.1021/acs.est.0c08712

Zhang, W., Morales, V. L., Cakmak, M. E., Salvucci, A. E., Geohring, L. D., Hay, A. G., et al. (2010). Colloid transport and retention in unsaturated porous media: effect of colloid input concentration. *Environ. Sci. Technol.* 44, 4965–4972. doi: 10.1021/es100272f

Zhong, H., Liu, G., Jiang, Y., Yang, J., Liu, Y., Yang, X., et al. (2017). Transport of bacteria in porous media and its enhancement by surfactants for bioaugmentation: a review. *Biotechnol. Adv.* 35, 490–504. doi: 10.1016/j.biotechadv.2017.03.009