Vietnam Journal of Earth Sciences, 46(4), 515-532, https://doi.org/10.15625/2615-9783/21133

Vietnam Journal of Earth Sciences
https://vjs.ac.vn/index.php/jse

Hybrid approach for permeability prediction in porous media: combining FFT simulations with machine learning

Hai-Bang Ly[1,*], Hoang-Long Nguyen[1], Viet-Hung Phan[2,3], Vincent Monchiet[3]

[1]*University of Transport Technology, Hanoi 100000, Vietnam*
[2]*University of Transport and Communications, Hanoi 100000, Vietnam*
[3]*Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, MSME UMR 8208, 77454, Marne-la-Vallée, France*

Received 25 May 2024; Received in revised form 15 June 2024; Accepted 13 July 2024

ABSTRACT

The prediction of permeability in porous media is a critical aspect in various scientific and engineering applications. This paper presents a machine learning (ML) model based on the XGBoost algorithm for predicting the permeability of porous media using microstructure characteristics. The seahorse optimization algorithm was employed to fine-tune the hyperparameters of the XGBoost algorithm, resulting in a model with predictive solid capabilities. Regression analysis and residual errors indicated that the model achieved good prediction results on the training and testing datasets, with RMSE values of 0.0494 and 0.0826, respectively. A SHAP value sensitivity analysis revealed that the essential inputs were the size of the inclusions, with the quantiles representing the maximum size of the inclusions being the most significant variables affecting permeability. The findings of this study have important implications for the design and optimization of porous media, and the XGBoost algorithm-based ML model provides a fast and accurate tool for predicting the permeability of porous media based on microstructure characteristics.

Keywords: Permeability, machine learning, porous media, FFT, optimization.

1. Introduction

Fluid flow simulations in porous media are important in various disciplines, with particular emphasis on geoscience and material science (Auriault and Boutin, 1994, 1993, 1992; de Borst, 2017). This computational process is pivotal in comprehending and forecasting an array of natural and industrial phenomena (de Borst, 2017), including groundwater dynamics, hydrocarbon recovery, carbon dioxide sequestration (Ewing, 1983; Huppert and Neufeld, 2014; Khalid Awan et al., 2015), and material behavior under diverse conditions. In geoscience, understanding the flow of fluids through porous media such as rocks and soils is indispensable for analyzing processes like hydrocarbon migration, contaminant transport, and groundwater movement (Bachu, 2008). Within material science, it is equally essential for designing and optimizing materials utilized in applications ranging from filtration and battery electrodes to fuel cells (Herzig et al., 1970), where fluid interaction with the material is of critical concern. Moreover, within the concrete field, the

*Corresponding author, Email: banglh@utt.edu.vn

formation of cracks in concrete is a significant concern that can significantly affect its permeability (Dietrich et al., 2005; Grassl, 2009). Concrete, a widely used construction material, is inherently porous, and its permeability is critical in determining its durability and service life (Li et al., 2019).

Permeability, a key parameter in these simulations, measures the ease with which a fluid can traverse porous media (Renard and De Marsily, 1997). The accurate prediction of permeability is of utmost importance, as it directly influences the reliability of flow simulations. For instance, an underestimation of the permeability of petroleum engineering could result in overestimating the energy required for oil or gas extraction. Conversely, overestimating permeability may lead to an excessively optimistic forecast of resource recovery (Sander et al., 2017). However, predicting permeability is complex due to the intricate microstructures inherent in porous media. These microstructures can exhibit substantial variability, even within a single specimen, and significantly influence fluid flow.

Consequently, accurately modeling these complex microstructures is a prerequisite for reliably predicting their permeability. This necessity presents a challenge, as the complex microstructures of porous media often necessitate high-resolution imaging and computationally demanding simulations (Borujeni et al., 2013). Furthermore, these microstructures may evolve due to erosion, deposition, or chemical reactions, thereby introducing an additional layer of complexity (Yasuhara and Elsworth, 2006).

Predicting permeability and fluid flow through real microstructures presents several challenges (Al-Omari and Masad, 2004). While analytical approaches can provide solutions for simple cases or regular microstructures, they often struggle with the complexity and irregularity of real-world microstructures (Monchiet et al., 2019; Wang, 2003, 2001). Furthermore, these approaches may require constrained boundary conditions that limit their applicability and accuracy. Finite Element Method (FEM) is a powerful numerical tool for simulating fluid flows, but it requires the creation of a mesh that accurately represents the complex microstructure (Burman and Hansbo, 2007; Correa and Loula, 2009; Ly et al., 2015). This process can be time-consuming and computationally intensive, particularly for large or intricate structures.

In some cases, meshing may even be practically impossible due to the extreme complexity of the microstructure. Lattice Boltzmann Method (LBM) (Pan et al., 2004) and Fast Fourier Transform (FFT)-based methods can handle complex microstructures without the need for meshing (Ly et al., 2016; Monchiet et al., 2009; Nguyen et al., 2013), making them attractive alternatives. However, these methods can be computationally expensive, particularly for large-scale simulations or when high accuracy is required (Ly et al., 2022). This can limit their practicality, especially when dealing with multiple samples or when real-time predictions are needed. Therefore, there is a need for an alternative approach that can accurately simulate permeability based on real microstructures while also focusing on computational efficiency. This could involve the development of new numerical methods, the optimization of existing methods to reduce computational costs, or the use of machine learning (ML) techniques to predict permeability based on microstructure data.

Over the past few decades, the field of civil engineering has witnessed a significant evolution in the application of Artificial Intelligence (AI) and ML (Phoon and Zhang, 2023; Phung et al., 2023; Thai, 2022; Vadyala et al., 2022). These technologies have transitioned from simple rule-based systems to sophisticated models capable of learning from complex data, driven by algorithm advances, increases in computational power, and the availability of large datasets. AI and ML have shown great potential in solving complex problems related to materials, structures, and geosciences (Ly et al., 2020; Nguyen et al., 2020; Van Phong et al., 2020; Xuan et al., 2024). One of the critical applications of AI and ML in civil engineering is the analysis of material behavior (Morgan and Jacobs, 2020; Wei et al., 2019). For instance, ML algorithms can be used to predict the mechanical properties of concrete based on its mix design and curing conditions. This can help optimize the mix design to achieve desired properties, reduce material waste, and improve the sustainability of concrete structures (Hasanipanah et al., 2023; Ly et al., 2021; Ly and Nguyen, 2024).

Similarly, ML can be used to predict the performance of geomaterials, such as soils and rocks (Nhu et al., 2023), under various loading conditions, which can aid in geotechnical engineering design. Another application of AI and ML in civil engineering is monitoring and assessing structures (Flah et al., 2021). ML algorithms can analyze sensor data from structures, such as bridges and buildings to detect signs of damage or deterioration (Zhou et al., 2017). This can help identify potential safety issues before they become critical and enable proactive maintenance. Regarding the problem of predicting permeability in porous media, AI and ML present a possible alternative to traditional simulation methods (Brunton et al., 2020; Ren et al., 2020). Instead of solving the fluid flow equations for a given microstructure, ML models can be trained to predict permeability directly from microstructure data (Erofeev et al., 2019; Phan and Ly, 2024; Tian et al., 2021). This can significantly reduce the computational cost, especially for complex microstructures. Moreover, once trained, ML models can make predictions almost instantaneously, enabling real-time analysis and decision-making. They can also handle uncertainties in the microstructure data, which can be challenging to account for in traditional simulations (Srinivasan et al., 2018). AI and ML offer promising avenues for solving complex problems in civil engineering, including permeability prediction in porous media. As these technologies evolve, they will likely play increasingly important roles in this field.

This study leverages AI and ML's power to enhance permeability prediction in porous media, a critical parameter in various civil engineering applications. The seahorse optimization algorithm (SHOA) is utilized to finely tune the hyperparameters of the XGBoost algorithm, a robust ML tool, which is then applied to a dataset generated from FFT simulations. The input space of the model encompasses the quantile distribution of the size and orientation of inclusions in the porous media. This study aims to optimize the ML model's performance to provide an accurate and efficient tool for predicting permeability in porous media, ultimately leading to improved design and management of civil engineering systems. Furthermore, a sensitivity analysis is conducted using SHAP values to evaluate the importance of each input feature on the predicted permeability, offering insights into the relationship between microstructural features and permeability.

2. Fast-Fourier Transform (FFT) simulation for fluid flow problem

2.1. Overview of FFT simulation

FFT-based simulation represents a computational approach utilized for the modeling and analysis of nonlinear composites (Michel et al., 1999; Moulinec and Suquet, 1998), then successfully applied in the context of predicting fluid flow and transport properties (Ly et al., 2016; Nguyen et al., 2013). This method leverages the FFT algorithm, a powerful tool for efficiently computing the discrete Fourier transform of a sequence, to provide a robust and accurate representation of the porous medium. In porous media simulation, FFT-based methods typically involve representing the porous medium as a binary image, where different pixel values denote solid and fluid phases. The image is then transformed into the frequency domain using the FFT algorithm, and the resulting data is used to solve the governing equations for fluid flow, such as the Stokes, Bingham, Darcy or Navier-Stokes equations. The solution is subsequently transformed back into the spatial domain to obtain the desired flow and transport properties.

FFT-based simulation offers several advantages for porous media modeling. It is highly efficient, particularly for large-scale simulations, and can easily handle complex geometries and microstructures. It also allows for directly calculating effective transport properties, such as permeability and diffusivity, without explicit pore-scale simulations (Mezhoud et al., 2020). However, FFT-based methods also have some limitations. They assume periodic boundary conditions that may not always be appropriate for representing real-world porous media. They also require regular grids, which can limit their ability to represent complex geometries and microstructures accurately. Despite these challenges, FFT-based simulation remains a valuable tool for porous media modeling, offering a powerful and efficient approach for predicting fluid flow and transport properties in complex porous media systems. Its robustness and accuracy make it a popular choice for researchers and engineers seeking to understand and optimize the behavior of porous media in a wide range of applications.

2.2. Permeability dataset

The present study utilized a dataset of permeability values to investigate ML algorithms' efficacy in predicting porous media permeability. Ly et al.'s paper elucidates the underlying principle of the problem to be solved. (Ly et al., 2022). A total of 2000 simulations were performed on generated microstructures, as illustrated in Fig. 1, where the inclusion and porous solid (matrix phase) are represented by yellow and blue colors, respectively. The microstructures were generated such that 100 inclusions were present in the unit cell, with the permeability of the porous phase set at 10^{-6}.

To introduce variability into the dataset, the positions of the 100 inclusions were randomly distributed inside the unit cell. The two dimensions along the Ox and Oy axes were randomly chosen to fall between 0.01 and 0.2, resulting in a diverse range of microstructures. An FFT simulation was conducted for each generated microstructure to obtain the corresponding permeability value. The resulting permeability values were then saved for subsequent use in ML simulations. This dataset provides a valuable resource for training and testing ML algorithms to accurately predict the permeability of porous media based on the characteristics of the microstructure. The dataset's diversity in terms of inclusion positions and dimensions enables the assessment of the ML algorithms' ability to generalize and make accurate predictions for a wide range of microstructures.

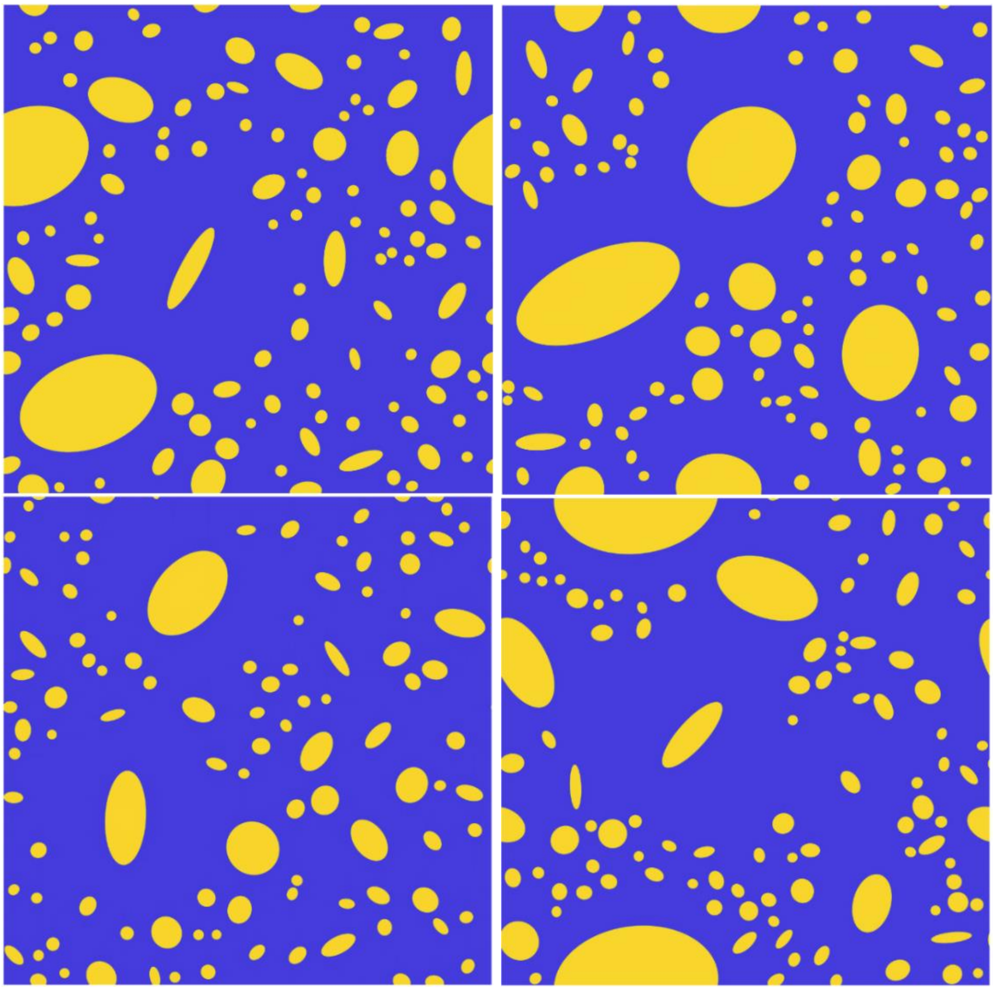

Figure 1. Examples of complex microstructure with 100 inclusions

2.3. Input space selection

In ML, selecting an appropriate input space is a critical factor that significantly influences the model's efficacy (Blum and Langley, 1997). The identification of relevant features that bear a direct relationship with the given problem is of paramount importance. In the context of the complex problem of fluid flow through generated 2D microstructures, a judicious choice of input features is essential to predict the permeability of a porous medium accurately.

First, the spatial arrangement of the 100 inclusions within the microstructure significantly affects the fluid flow patterns and, consequently, the permeability of the porous medium. Therefore, the positions of the inclusions constitute a critical input feature for the ML model.

Second, the dimensions of the inclusions can influence the porous medium's fluid flow and transport properties. Thus, the sizes of the 100 inclusions represent another important input feature of the ML model.

Third, the permeability of the porous phase itself can significantly impact the porous medium's overall fluid flow and transport properties. Therefore, including this parameter in the input space can enhance the accuracy of the ML model.

Finally, the orientation of the inclusions can affect the fluid flow patterns and, consequently, the permeability of the porous medium. Thus, the angle of rotation of each inclusion represents a critical input feature of the ML model.

In summary, the input space for the ML model should encompass the positions, sizes, and angles of rotation of the 100 inclusions and the permeability of the porous phase. These features capture the microstructure characteristics that influence the porous medium's fluid flow and transport properties. By considering these factors, the ML model can accurately predict the permeability of a porous medium based on the microstructure characteristics, thereby providing insights into the behavior of complex porous media systems. In the context of ML simulations, incorporating all information of the 100 inclusions into the input space would result in a considerable input space, leading to significant computational challenges. Therefore, based on a preliminary analysis, a judicious selection of inputs was made to ensure computational feasibility while maintaining the model's accuracy.

3. ML methods

3.1. Extreme Gradient Boosting

XGBoost, short for Extreme Gradient Boosting, is a powerful and versatile ML algorithm that has gained significant popularity in predictive modeling (Chen and Guestrin, 2016). It implements gradient boosting, a technique that builds a stage-wise ensemble of weak prediction models, typically decision trees, to create a robust predictive model.

The primary working mechanism of XGBoost involves iteratively adding new models to the ensemble, with each subsequent model designed to correct the errors made by the previous ones. This is achieved by having each new model focus on the residuals, or the differences between the actual and predicted values, of the previous model. By iteratively refining the predictions in this manner, XGBoost can achieve high accuracy and robustness.

XGBoost offers several advantages, making it a popular choice for many predictive modeling tasks. Its use of gradient boosting allows it to achieve high accuracy and handle complex data sets. It also includes several regularization techniques to prevent overfitting, such as L1 and L2 regularization of the weights of the decision trees, and a method called shrinkage, which scales down the contributions of each tree to the final prediction. Furthermore, XGBoost supports parallel processing, which can significantly speed up computation time and handle missing values in the data.

However, XGBoost also has some potential drawbacks. Its high complexity makes it more difficult to interpret and understand than simpler models. It also has several hyperparameters that need to be tuned, which can be time-consuming and require expert knowledge. Moreover, XGBoost can be resource-intensive, mainly when working with large data sets, which can be a limitation in some applications. Despite these challenges, XGBoost remains a relevant tool in the ML toolbox, offering a powerful and flexible solution for many predictive modeling problems.

3.2. Seahorse Optimizer Algorithm (SHOA)

The SHOA is a sophisticated, nature-inspired algorithm for complex optimization problems (Zhao et al., 2023). It derives its principles from the distinctive foraging behavior and movement patterns exhibited by seahorses adept at navigating and hunting within intricate marine environments. The algorithm initiates with a population of seahorses (pop_size), each symbolizing a potential solution within the problem space. It emulates the seahorse's hunting strategy, where they camouflage by changing their skin color, maneuvering their heads and tails independently, and utilizing their elongated snouts to capture prey. This strategy is translated into a series of mathematical operations within the algorithm, enabling each seahorse to update its position based on its own and other seahorses' experiences. The SHOA maintains a balance between exploration (scanning the entire space for potential solutions) and exploitation (concentrating on promising areas) to efficiently locate the optimal solution.

The SHOA offers several advantages, including versatility, efficiency, and robustness. Its applicability extends to various optimization problems, such as engineering design, feature selection, and parameter tuning. SHOA demonstrates resilience against local optima, enabling it to discover the global optimum even in complex problem spaces frequently. However, SHOA faces challenges related to parameter selection, such as determining the optimal number of seahorses and iterations through experimentation. Its complex nature, involving extensive mathematical operations, can complicate implementation and comprehension compared to simpler algorithms. Furthermore, while effective for various problems, its performance on large-scale or high-dimensional tasks requires further exploration.

Overall, SHOA was chosen in this study for hyperparameter tuning due to its effective balance between global exploration and local exploitation, which is crucial for avoiding local minima and achieving optimal solutions efficiently.

3.3. Cross-validation

Cross-validation (CV) is a crucial technique in evaluating the performance of ML models. In this study, a 5-fold CV is employed to optimize the hyperparameters of the XGBoost algorithm (Fig. 2). This approach involves dividing the dataset into five equal subsets, where each subset is used as a validation set. In contrast, the remaining four are used for training. By iterative training and validating the model on different subsets of the data, 5-fold CV provides a robust and reliable estimate of the model's performance, allowing for selecting the optimal hyperparameters and improving the model's predictive accuracy.

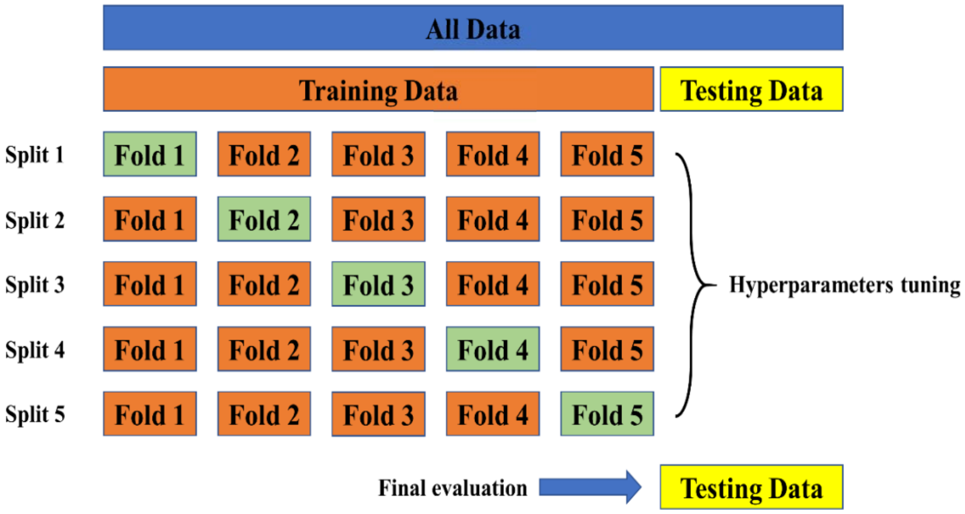

Figure 2. 5-fold CV illustration

3.4. Model metrics

The root mean square error (RMSE), mean absolute percentage error (MAPE), correlation coefficient (R), and mean absolute error (MAE) are metrics used to evaluate the performance of the predictive models in this study. RMSE is a standard metric that calculates the square root of the average of squared differences between predicted and actual values, making it highly sensitive to significant errors. R is a statistical metric that measures the strength and direction of the linear relationship between two variables, measuring how well the observed outcomes correlate with the predicted outcomes. Lastly, MAE measures the average magnitude of the errors in a set of predictions without considering their direction, making it less sensitive to significant errors than RMSE. These metrics offer different perspectives on the performance of a predictive model, and the choice of which metric to use depends on the specific goals and requirements of the modeling task at hand. These metrics are formulated in the relevant literature (Ly and Nguyen, 2024; Phan and Ly, 2024).

4. Results and discussions

4.1. Feature selection and database analysis

It is necessary to establish several notations utilized in this study to facilitate understanding. Fig. 3 illustrates the relevant parameters, where the major and minor axes of the elliptic inclusion are represented by a and b, respectively. The angle of rotation of the inclusion is defined by the intersection of the Ox and the ellipse's major axis. Furthermore, the position of the elliptic inclusion is specified by its coordinates in the Ox and Oy directions.

Referring to Fig. 4, the distribution of the primary characteristics of the 100 inclusions is presented, including the sizes of the major (a) and minor (b) axes, their respective positions along the Ox and Oy axes, and the rotation angle (Phi). It can be observed that the sizes of the inclusions are predominantly small, with over 95% of the inclusions having a size of less than 0.06. Only a few inclusions have a size larger than 0.06. In terms of the position of the inclusions, they appear to be evenly distributed throughout the unit cell. A similar trend is observed for the rotation angle, with the values of Phi being relatively uniformly distributed from 0 to approximately 180 degrees.

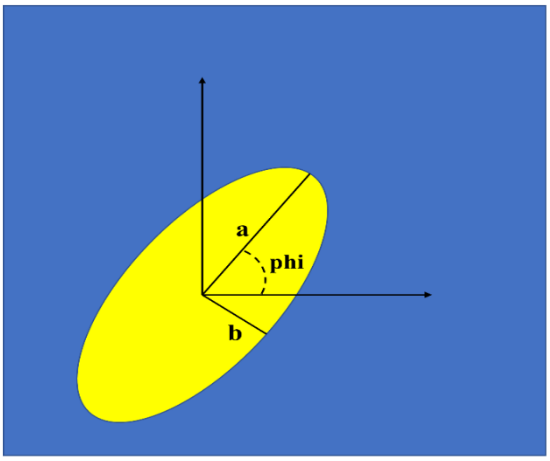

Figure 3. Illustration of the notation of the porous media in this study

In addressing the problem, the most straightforward approach (ML) would be to consider all inclusions as input parameters. However, this method could significantly increase the input space, potentially resulting in more than 100 features. To simplify the process, the authors propose to extract only relevant information from the input space and perform ML simulations on this reduced dataset. Four cases are considered for initial assessment, as presented in Table 1. Different quantile levels are taken to retrieve significant information related to the microstructure for ML simulation. The ML modeling uses the XGBoost algorithm with default hyperparameters for simplification. The results regarding the training, testing, and overall dataset are presented in Table 2.

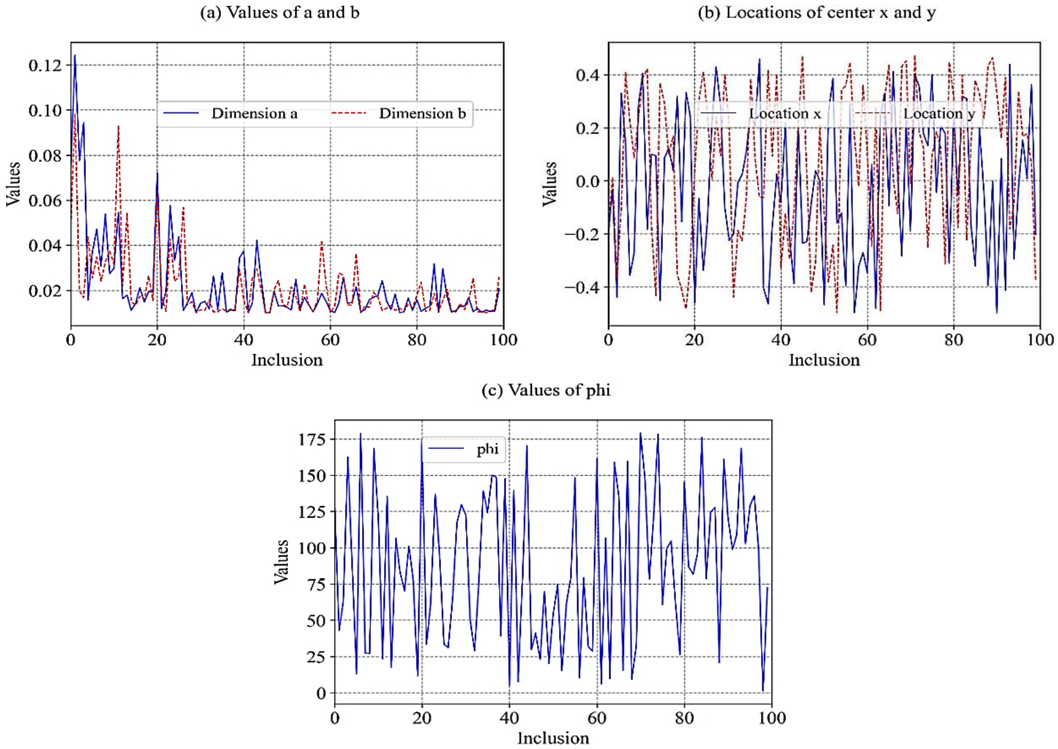

Figure 4. The distribution of the primary characteristics of the 100 inclusions

Table 1. Cases were considered for the initial assessment of feature selection

Cases	Quantile levels considered
1	0; 0.25; 0.5; 0.75; 1
2	0; 0.2; 0.4; 0.6; 0.8; 1
3	0; 0.5; 0.6; 0.7; 0.8; 0.9; 1
4	0; 0.5; 0.8; 0.85; 0.9; 0.95; 1

Table 2. Permeability dataset and statistical analysis

Case	Data	RMSE	MAE	R	MAPE
1	Training	0.004	0.003	0.999	0.002
	Testing	0.106	0.078	0.657	0.049
	All	0.058	0.025	0.893	0.016
2	Training	0.003	0.002	0.999	0.001
	Testing	0.105	0.076	0.664	0.047
	All	0.057	0.024	0.895	0.015
3	Training	0.003	0.002	0.999	0.001
	Testing	0.099	0.071	0.701	0.044
	All	0.054	0.023	0.906	0.014
4	Training	0.003	0.002	0.999	0.001
	Testing	0.099	0.071	0.709	0.044
	All	0.054	0.023	0.908	0.014

The results of the study indicate that the accuracy of the ML model in predicting permeability increases with the inclusion of higher quantile levels of inclusion sizes. In case 1, where only five levels were considered, covering the entire range from 0 to 100%, the accuracy of the ML model was relatively low, with an R-value of 0.657. However, as more significant quantile levels were included in cases 2 and 3, the R values increased to 0.664 and 0.701, respectively. The highest accuracy was achieved in case 4, where the quantile levels focused on the larger inclusion sizes, greater than 80%. In this case, the ML model's accuracy was acceptable, with an R-value of 0.709. Therefore, the input space for this problem should include the features considered in case 4.

As mentioned earlier, the quantile levels of the size of inclusion "a" and "b" were set at 0, 0.5, 0.8, 0.85, 0.9, 0.95, and 1, and seven inputs related to each quantile level were chosen. Additionally, seven inputs related to the quantile distribution of the orientation of the inclusions were also included. As a preliminary study aimed at predicting the permeability of image-based microstructures, a simplification was made by assuming a constant permeability of the porous phase. In other words, all the values of the permeability of the porous phase were taken as 10^{-6}.

Overall, the problem to be solved comprises 21 inputs, including the quantile distributions of the sizes and orientations of the most significant inclusions, and one output, which is the predicted permeability of the porous medium (Table 3). This simplification enables the ML model to accurately predict the permeability of the porous medium while maintaining computational feasibility, thereby providing insights into the behavior of complex porous media systems.

Table 3. Permeability dataset and statistical analysis

Variable	Average	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	max
X_1	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
X_2	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
X_3	0.03	0.00	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03
X_4	0.03	0.00	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04
X_5	0.04	0.00	0.02	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.06
X_6	0.05	0.01	0.03	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.10
X_7	0.12	0.03	0.05	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.20
X_8	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
X_9	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
X_{10}	0.03	0.00	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.04
X_{11}	0.03	0.00	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04
X_{12}	0.04	0.00	0.02	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.05
X_{13}	0.05	0.01	0.03	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.09
X_{14}	0.12	0.03	0.05	0.08	0.09	0.10	0.11	0.12	0.12	0.13	0.15	0.16	0.20
X_{15}	1.78	1.76	0.00	0.18	0.41	0.66	0.92	1.25	1.62	2.11	2.84	4.12	15.44
X_{16}	90.17	8.71	63.93	79.03	82.94	85.51	88.00	90.08	92.32	94.77	97.75	101.29	119.87
X_{17}	143.67	7.08	117.27	134.74	137.52	139.93	142.06	144.05	145.66	147.65	149.67	152.49	167.71
X_{18}	152.56	6.30	124.39	144.52	147.14	149.36	151.14	152.99	154.54	156.24	158.07	160.36	170.76
X_{19}	161.47	5.21	140.30	154.74	157.19	158.82	160.54	161.75	163.01	164.41	166.00	168.09	174.34
X_{20}	170.38	3.87	150.36	165.20	167.38	168.76	169.84	170.80	171.81	172.72	173.70	174.88	178.23
X_{21}	178.24	1.74	163.44	176.10	177.16	177.83	178.41	178.79	179.09	179.40	179.61	179.79	180.00
Y	1.56	0.13	1.30	1.42	1.46	1.49	1.52	1.55	1.57	1.61	1.66	1.73	2.50

4.2. Hyperparameter selection

This section describes the use of the SHOA to tune the hyperparameters of the XGBoost algorithm finely is described. 70% of the data (training set) is utilized to accomplish this. CV with 5 folds is applied to the training dataset, and the SHOA is employed to define four critical hyperparameters of XGBoost, namely n_estimator, learning_rate, max_depth, and subsamples. The n_estimator parameter determines the number of trees in the gradient boosting ensemble, while the learning_rate parameter controls the step size in the gradient descent algorithm used to minimize the loss function. The max_depth parameter specifies the maximum depth of each tree in the ensemble, and the subsamples parameter defines the fraction of samples to be used for training each tree. The pop_size of the SHOA is chosen as 30 to ensure a balance between computing time and effectiveness, and the number of iterations is chosen as 200. For each iteration, the 5-fold CV score is estimated, and the best set of hyperparameters is defined as the one that gives the lowest RMSE CV score between the ML-based and

FFT-based permeabilities. This approach leverages the SHOA's power to optimize the XGBoost algorithm's hyperparameters, thereby improving the model's accuracy. Using CV and minimizing the RMSE CV score can identify the optimal set of hyperparameters, resulting in a more accurate and reliable predictive model.

As can be observed (Fig. 5), the CV score curve tends to reach lower values after several iterations, indicating that SHOA is effectively fine-tuning the hyperparameters of the XGBoost algorithm. The RMSE is 0.068 from the first iteration, and it decreases to 0.067 and 0.0665 after 200 iterations, demonstrating a gradual improvement in the model's accuracy. Notably, after only 25 iterations, the SHOA reaches its most effective values in finely tuning the hyperparameters of XGBoost. This suggests that the SHOA is a computationally efficient optimization algorithm that can quickly identify the optimal set of hyperparameters for the XGBoost algorithm. At convergence, the identified best set of hyperparameters is presented in Table 4.

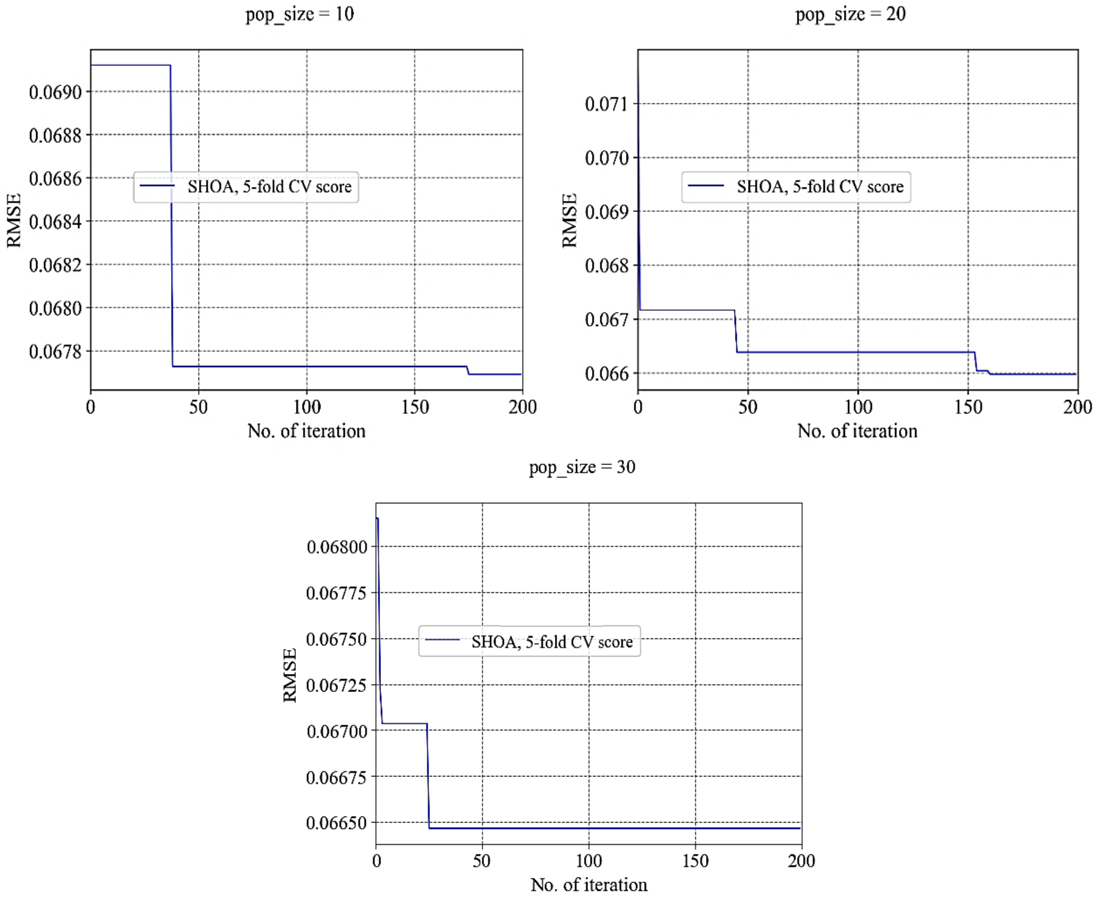

Figure 5. Fine-tuning process of XGBoost using SHOA with pop_size = 10, 20, and 30 over 200 iterations

Table 4. XGBoost hyperparameters selected for fine-tuning by SHOA

	n_estimator	**learning_rate**	**max_depth**	**subsamples**
Min	5	0.001	1	0.1
Max	1000	0.7	16	1
Best value by SHOA	180	0.020	4	0.150

For comparison purposes, the computation time per iteration is shown in Fig. 6. Specifically, the tuning process involved 200 iterations, and the computation time varied with the pop_size used. For a pop_size of 10, the hyperparameter tuning took approximately 8 s per iteration. Increasing the pop_size to 20 resulted in a computation time of about 18 s per iteration. With a pop_size of 30, the computation time was around 100 s per iteration, totaling around 5.5 hours for the entire process. The objective function, RMSE of the 5-fold CV score, is 0.0665 at convergence. This indicates that the SHOA can effectively optimize the hyperparameters of the XGBoost algorithm to achieve a high level of accuracy in predicting the permeability of porous media based on the microstructure characteristics.

In summary, using the SHOA to optimize the hyperparameters of the XGBoost algorithm is a computationally efficient and effective approach for predicting the permeability of porous media. The optimal set of hyperparameters can be identified by minimizing the RMSE of the 5-fold CV score, resulting in a more accurate and reliable predictive model.

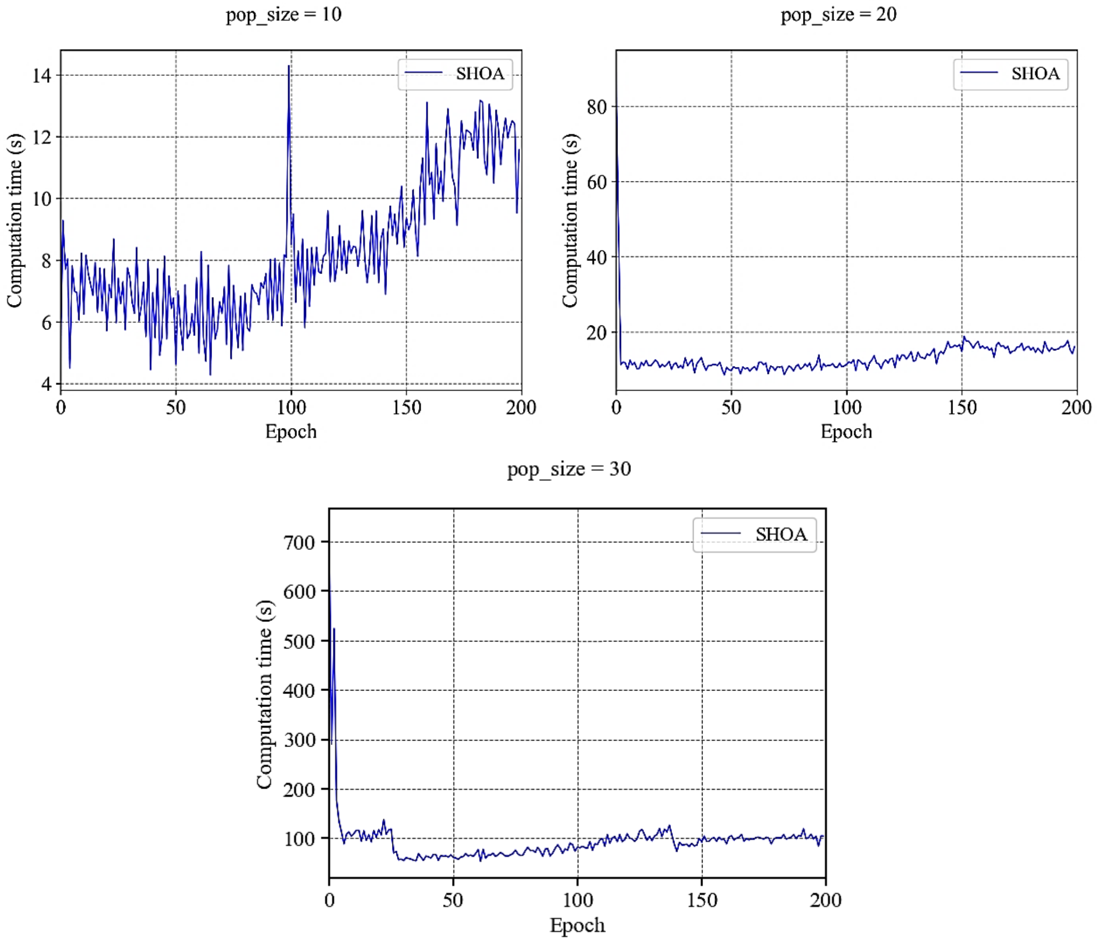

Figure 6. Computation time per epoch of XGBoost fine-tuning process

4.3. Learning curve of the algorithm

Overfitting is a critical issue in ML, and it needs to be carefully addressed and checked once the model is constructed. One of the many ways to effectively detect overfitting is using the learning curve. The learning curve of the developed XGBoost model is plotted in Fig. 7, showing the model's performance as the size of the training set increases.

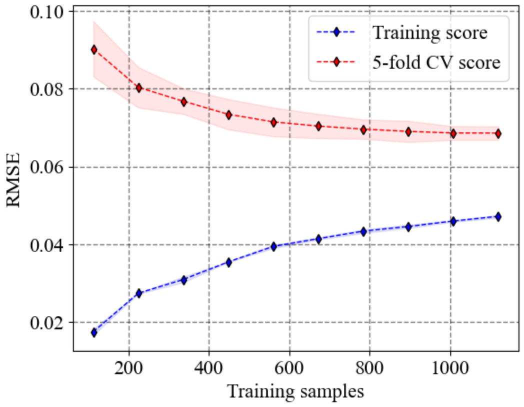

Figure 7. Learning curve of the XGBoost model

The learning curve analysis reveals how the model behaves with different training set sizes and 5-fold CV. It shows that increasing the size of the training set and utilizing 70% of the data enhances the model's ability to generalize to new data. The decreasing CV score as the training samples grow indicates improved generalization capability. At the same time, the rising trend in the training score suggests effective learning without overfitting the training data. These findings demonstrate that utilizing 70% of the dataset further enhances the model's performance, effectively minimizing overfitting. Thus, optimizing XGBoost hyperparameters with SHOA proves to be an effective strategy for predicting porous media permeability with robust and reliable outcomes.

4.4. Model predictive ability

This section describes the model's predictive ability by presenting the regression analysis and residual errors for the training and testing datasets. The regression plots are shown in Fig. 8, along with the respective histograms of the distribution of values. For the training dataset, it can be observed that the model achieves good prediction results where the training data points are close to the diagonal line. This indicates that the model accurately captures the relationship between the input features and the output variable. Several errors are found for the testing dataset, but the overall trend is good, indicating that the model can generalize well to new data. The computed RMSE values are 0.0494 and 0.0826, the R values are 0.916 and 0.808, the MAPE values are 0.0240 and 0.0346, and the MAE values are 0.0378 and 0.0564 for the training and testing sets, respectively.

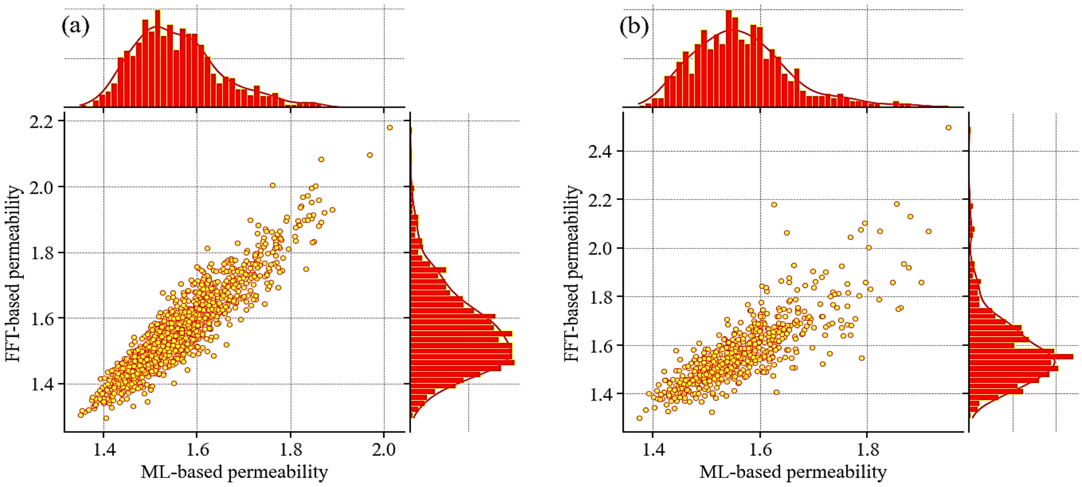

Figure 8. Performance of the XGBoost model in predicting the permeability of porous media (a) training and (b) testing

These results demonstrate the model has good predictive ability, with low errors and high correlation coefficients for both the training and testing datasets. Overall, the model shows promising results and can be a useful tool for various applications in predicting the permeability of porous media.

4.5. Sensitivity analysis

In this section, a SHAP value sensitivity analysis is conducted to evaluate the influence of inputs on permeability (Fig. 9). The analysis is based on two viewpoints using bee-swarm SHAP analysis, namely the mean value of SHAP and the maximum value of SHAP. Among the 21 inputs considered in the database, only the 9 most important ones are shown, while the remaining inputs are shown as a sum effect.

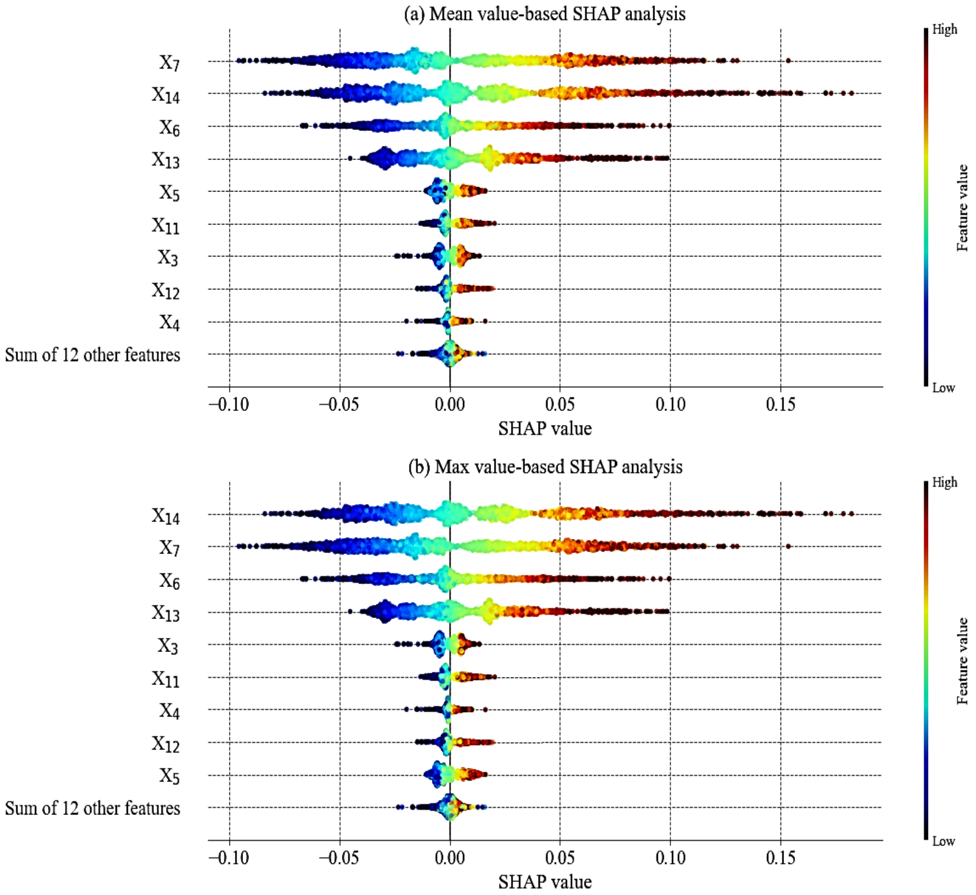

Figure 9. Shap values analysis of XGBoost model using bee-swarm plots: (a) based on mean SHAP values, and (b) based on max SHAP values

Interestingly, the most essential input is the size of the inclusions, ranging from the most prominent sizes to smaller ones. X_7 and X_{14}, which represent the maximum size of the inclusion, are classified as the most important variables affecting the permeability. X_6 and X_{13} represent the quantile at 95% of the size distribution and are also identified as essential variables. The following vital variables are the smaller sizes, representing the quantiles at

90%, 85%, and 80%. On the other hand, the orientation of the inclusion does not significantly affect the permeability of the porous medium in this case, or at least they are less influential than the dimensions. This suggests that the size of the inclusions is the dominant factor in determining the permeability of the porous medium, while the orientation of the inclusions has a minimal effect. The SHAP value sensitivity analysis provides helpful information on the relative importance of the input variables and their effects on the permeability of the porous medium. This information can be used to optimize porous media design and improve the accuracy of predictive models.

4.6. Discussions

The findings of this study hold significant implications for predicting permeability in porous media based on microstructure characteristics. The selection of the XGBoost algorithm was driven by its robustness and scalability in capturing intricate feature relationships through gradient boosting and effective regularization methods to prevent overfitting. Its scalability enables efficient handling of large datasets and complex feature spaces typical in porous media simulations. Moreover, utilizing the SHOA for hyperparameter tuning markedly improved model performance. These advantages position XGBoost as the preferred algorithm for our study's objectives.

However, the study acknowledges limitations, such as simplifying permeability variations within the porous phase, which may not fully reflect real-world complexities. Additionally, the model's sensitivity to input data quality and quantity highlights ongoing challenges in deploying robust models.

The study employed learning curve analysis to address overfitting, demonstrating the model's ability to generalize without merely memorizing training data patterns. The SHAP value sensitivity analysis highlighted the importance of inclusion size, where larger sizes and higher quantile levels notably influence permeability predictions. Enhancing the granularity of quantile levels in the input space, particularly for larger inclusion sizes, could further enhance model accuracy.

Future research directions should consider expanding datasets to encompass diverse microstructure characteristics and realistic permeability variations across various porous media types. Exploring alternative optimization algorithms for hyperparameter tuning could also provide additional insights into optimizing model performance under diverse conditions.

5. Conclusions

The study presented an XGBoost algorithm-based ML model for predicting the permeability of porous media based on microstructure characteristics. The SHOA was utilized to finely tune the hyperparameters of the XGBoost algorithm, resulting in a model with good predictive ability. The regression analysis and residual errors showed that the model achieved good prediction results for the training and testing datasets, with RMSE values of 0.0494 and 0.0826, respectively. The SHAP value sensitivity analysis revealed that the most important input was the size of the inclusions, with the quantiles representing the maximum size of the inclusion being the most important variables affecting the permeability.

The findings of this study have important implications for the design and optimization of porous media. The XGBoost algorithm-based ML model provides a fast and accurate tool for predicting the permeability of porous media based on microstructure characteristics, which can aid in designing and optimizing porous media for various applications. Furthermore, the SHAP value sensitivity analysis provides insights into the relationship between the input variables and the permeability of the porous medium, which can guide the selection of input variables for future studies.

Acknowledgments

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.03-2019.23.

References

Al-Omari A., Masad E., 2004. Three dimensional simulation of fluid flow in X-ray CT images of porous media. Num Anal Meth Geomechanics, 28, 1327–1360. https://doi.org/10.1002/nag.389.

Auriault J.L., Boutin C., 1992. Deformable porous media with double porosity. Quasi-statics. I: Coupling effects. Transport in Porous Media, 7, 63–82.

Auriault J.L., Boutin C., 1993. Deformable porous media with double porosity. Quasi-statics. II: Memory effects. Transport in Porous Media, 10, 153–169.

Auriault J.L., Boutin C., 1994. Deformable porous media with double porosity III: Acoustics. Transport in Porous Media, 14, 143–162.

Bachu S., 2008. CO_2 storage in geological media: Role, means, status and barriers to deployment. Progress in Energy and Combustion Science, 34, 254–273. https://doi.org/10.1016/j.pecs.2007.10.001.

Blum A.L., Langley P., 1997. Selection of relevant features and examples in machine learning. Artificial intelligence 97, 245–271.

Borujeni A.T., Lane N.M., Thompson K., Tyagi M., 2013. Effects of image resolution and numerical resolution on computed permeability of consolidated packing using LB and FEM pore-scale simulations. Computers & Fluids, 88, 753–763.

Brunton S.L., Noack B.R., Koumoutsakos P., 2020. Machine Learning for Fluid Mechanics. Annu. Rev. Fluid Mech, 52, 477–508. https://doi.org/10.1146/annurev-fluid-010719-060214.

Burman E., Hansbo P., 2007. A unified stabilized method for Stokes' and Darcy's equations. Journal of Computational and Applied Mathematics, 198, 35–51.

Chen T., Guestrin C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–794.

Correa M.R., Loula A.F.D., 2009. A unified mixed formulation naturally coupling Stokes and Darcy flows. Computer Methods in Applied Mechanics and Engineering, 198, 2710–2722.

de Borst R., 2017. Fluid flow in fractured and fracturing porous media: A unified view. Mechanics Research Communications, Multi-Physics of Solids at Fracture, 80, 47–57. https://doi.org/10.1016/j.mechrescom.2016.05.004.

Dietrich P., Helmig R., Sauter M., Hötzl H., Köngeter J., Teutsch G., 2005. Flow and transport in fractured porous media. Springer Berlin, Heidelberg. http://doi.org/10.1007/b138453.

Erofeev A., Orlov D., Ryzhov A., Koroteev D., 2019. Prediction of Porosity and Permeability Alteration Based on Machine Learning Algorithms. Transp Porous Med, 128, 677–700. https://doi.org/10.1007/s11242-019-01265-3.

Ewing R.E., 1983. 1. Problems Arising in the Modeling of Processes for Hydrocarbon Recovery, in: Ewing, R.E. (Ed.), The Mathematics of Reservoir Simulation. Society for Industrial and Applied Mathematics, 3–34. https://doi.org/10.1137/1.9781611971071.ch1.

Flah M., Nunez I., Ben Chaabene W., Nehdi M.L., 2021. Machine learning algorithms in civil structural health monitoring: a systematic review. Archives of Computational Methods in Engineering, 28, 2621–2643.

Grassl P., 2009. A lattice approach to model flow in cracked concrete. Cement and Concrete Composites, 31, 454–460.

Hasanipanah M., Abdullah R.A., Iqbal M., Ly H.-B., 2023. Predicting Rubberized Concrete Compressive Strength Using Machine Learning: A Feature Importance and Partial Dependence Analysis. Journal of Science and Transport Technology, 3, 27–44.

Herzig J.P., Leclerc D.M., Goff P.Le., 1970. Flow of Suspensions through Porous Media Application to Deep Filtration. Industrial & Engineering Chemistry, 62, 8–35. https://doi.org/10.1021/ie50725a003.

Huppert H.E., Neufeld J.A., 2014. The Fluid Mechanics of Carbon Dioxide Sequestration. Annu. Rev. Fluid Mech, 46, 255–272. https://doi.org/10.1146/annurev-fluid-011212-140627.

Khalid Awan U., Tischbein B., Martius C., 2015. Simulating Groundwater Dynamics Using Feflow-3D Groundwater Model Under Complex Irrigation and Drainage Network of Dryland Ecosystems of Central Asia. Irrigation and Drainage, 64, 283–296. https://doi.org/10.1002/ird.1897.

Li X., Li D., Xu Y., 2019. Modeling the effects of microcracks on water permeability of concrete using 3D discrete crack network. Composite Structures, 210, 262–273.

Ly H.-B., Asteris P.G., Pham T.B., 2020. Accuracy assessment of extreme learning machine in predicting soil compression coefficient. Vietnam Journal of Earth Sciences, 42(3), 228–336. https://doi.org/10.15625/0866-7187/42/3/14999.

Ly H.-B., Le Droumaguet B., Monchiet V., Grande D., 2015. Designing and modeling doubly porous polymeric materials. The European Physical Journal Special Topics, 224, 1689–1706.

Ly H.B., Monchiet V., Grande D., 2016. Computation of permeability with Fast Fourier Transform from 3-D digital images of porous microstructures. International Journal of Numerical Methods for Heat & Fluid Flow 26(5), 1328–1345.

Ly H.-B., Nguyen T.-A., 2024. Machine learning-driven innovations in green eco-environmental rubberized concrete design towards sustainability. Materials Today Communications, 39, 108551.

Ly H.-B., Nguyen T.-A., Tran V.Q., 2021. Development of deep neural network model to predict the compressive strength of rubber concrete. Construction and Building Materials 301, 124081.

Ly H.-B., Phan V.-H., Monchiet V., Nguyen H.-L., Nguyen-Ngoc L., 2022. Numerical investigation of macroscopic permeability of biporous solids with elliptic vugs. Theoretical and Computational Fluid Dynamics, 36, 689–704.

Mezhoud S., Monchiet V., Bornert M., Grande D., 2020. Computation of macroscopic permeability of doubly porous media with FFT based numerical homogenization method. European Journal of Mechanics-B/Fluids, 83, 141–155.

Michel J.-C., Moulinec H., Suquet P., 1999. Effective properties of composite materials with periodic microstructure: a computational approach. Computer Methods in Applied Mechanics and Engineering, 172, 109–143.

Monchiet V., Bonnet G., Lauriat G., 2009. A FFT-based method to compute the permeability induced by a Stokes slip flow through a porous medium. Comptes Rendus Mécanique, 337, 192–197.

Monchiet V., Ly H.-B., Grande D., 2019. Macroscopic permeability of doubly porous materials with cylindrical and spherical macropores. Meccanica, 54, 1583–1596.

Morgan D., Jacobs R., 2020. Opportunities and Challenges for Machine Learning in Materials Science. Annu. Rev. Mater. Res., 50, 71–103. https://doi.org/10.1146/annurev-matsci-070218-010015.

Moulinec H., Suquet P., 1998. A numerical method for computing the overall response of nonlinear composites with complex microstructure. Computer Methods in Applied Echanics and Engineering, 157, 69–94.

Nguyen T.-A., Ly H.-B., Jaafario A., Pham B.T., 2020. Estimation offriction capacity of driven piles in clay using. Vietnam Journal of Earth Sciences, 42(3), 265–275. https://doi.org/10.15625/0866-7187/42/3/15182.

Nguyen T.-K., Monchiet V., Bonnet G., 2013. A Fourier based numerical method for computing the dynamic permeability of periodic porous media. European Journal of Mechanics-B/Fluids, 37, 90–98.

Nhu V.-H., Thai B.P., Tien D.B., 2023. A novel swarm intelligence optimized extreme learning machine for predicting soil shear strength: A case study at Hoa Vuong new urban project (Vietnam). Vietnam Journal of Earth Sciences, 45(2), 219–237. https://doi.org/10.15625/2615-9783/18338.

Pan C., Hilpert M., Miller C.T., 2004. Lattice-Boltzmann simulation of two-phase flow in porous media. Water Resources Research, 40, W01501. Doi: 10.1029/2003WR002120.

Phan V.-H., Ly H.-B., 2024. RIME-RF-RIME: A novel machine learning approach with SHAP analysis for predicting macroscopic permeability of porous media. Journal of Science and Transport Technology, 58–71.

Phoon K.-K., Zhang W., 2023. Future of machine learning in geotechnics. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 17, 7–22. https://doi.org/10.1080/17499518.2022.2087884.

Phung B.-N., Le T.-H., Nguyen M.-K., Nguyen T.-A., Ly H.-B., 2023. Practical Numerical Tool for Marshall Stability Prediction Based On Machine Learning: An Application for Asphalt Concrete Containing Basalt Fiber. Journal of Science and Transport Technology, 27–45.

Ren F., Hu H., Tang H., 2020. Active flow control using machine learning: A brief review. J Hydrodyn, 32, 247–253. https://doi.org/10.1007/s42241-020-0026-0.

Renard P., De Marsily G., 1997. Calculating equivalent permeability: a review. Advances in Water Resources, 20, 253–278.

Sander R., Pan Z., Connell L.D., 2017. Laboratory measurement of low permeability unconventional gas reservoir rocks: A review of experimental methods. Journal of Natural Gas Science and Engineering, 37, 248–279.

Srinivasan G., Hyman J.D., Osthus D.A., Moore B.A., O'Malley D., Karra S., Rougier E., Hagberg A.A., Hunter A., Viswanathan H.S., 2018. Quantifying topological uncertainty in fractured systems using graph theory and machine learning. Scientific Reports, 8, 11665.

Thai H.-T., 2022. Machine learning for structural engineering: A state-of-the-art review, in: Structures. Elsevier, 448–491.

Tian J., Qi C., Sun Y., Yaseen Z.M., Pham B.T., 2021. Permeability prediction of porous media using a combination of computational fluid dynamics and hybrid machine learning methods. Engineering with Computers, 37, 3455–3471. https://doi.org/10.1007/s00366-020-01012-z.

Vadyala S.R., Betgeri S.N., Matthews J.C., Matthews E., 2022. A review of physics-based machine learning in civil engineering. Results in Engineering, 13, 100316.

Van Phong T., Ly H.-B., Trinh P.T., Prakash I., BTJVJOES P., 2020. Landslide susceptibility mapping using Forest by Penalizing Attributes (FPA) algorithm based machine learning approach. Vietnam J. Earth Sci., 42(3), 237–246. https://doi.org/10.15625/0866-7187/42/3/15047.

Wang C.Y., 2001. Stokes flow through a rectangular array of circular cylinders. Fluid Dynamics Research, 29, 65–80. https://doi.org/10.1016/S0169-5983(01)00013-2.

Wang C.Y., 2003. Stokes slip flow through square and triangular arrays of circular cylinders. Fluid Dynamics Research, 32, 233–246. https://doi.org/10.1016/S0169-5983(03)00049-2.

Wei J., Chu X., Sun X., Xu K., Deng H., Chen J., Wei Z., Lei M., 2019. Machine learning in materials science. InfoMat, 1, 338–358. https://doi.org/10.1002/inf2.12028.

Xuan B.T., Thuy D.L., Van P.T., Hong N.V., Van Le H., Nguyen D.D., Prakash I., Thanh T.P., Thai B.B., 2024. Groundwater potential zoning using Logistics Model Trees based novel ensemble machine learning model. Vietnam Journal of Earth Sciences, 46(2), 272–281. https://doi.org/10.15625/2615-9783/20316.

Yasuhara H., Elsworth D., 2006. A numerical model simulating reactive transport and evolution of fracture permeability. Int. J. Numer. Anal. Meth. Geomech, 30, 1039–1062. https://doi.org/10.1002/nag.513.

Zhao S., Zhang T., Ma S., Wang M., 2023. Seahorse optimizer: a novel nature-inspired meta-heuristic for global optimization problems. Appl Intell, 53, 11833–11860. https://doi.org/10.1007/s10489-022-03994-3.

Zhou H., Liang X., Wang Z., Zhang X., Xing F., 2017. Bond deterioration of corroded steel in two different concrete mixes. Struct. Eng. Mech, 63, 725–734.