Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering) in press
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.A2500127

Predicting permeability coefficients of earth-rock material using improved GAN and explainable ensemble learning under small sample conditions

Chengyu YU, Hongling YU✉, Xiaofeng QU, Baoxi LIU, Liangsi XU, Xinyu LIU, Xiangyu CHEN

College of Water Resources and Civil Engineering, China Agricultural University, Beijing 100083, China

Abstract: Accurate prediction of the permeability coefficient is crucial for evaluating the compaction quality of earthworks. However, during the compaction process, on-site testing is often time-consuming and expensive, leading to fewer samples, which affects prediction accuracy. Moreover, most current predictive models have limited capabilities and tend to be black-box models with poor explainability. To overcome these issues, in this study we propose a new method to predict the permeability coefficient of earth-rock material, based on an improved Generative Adversarial Network (GAN) and explainable OOA-HL-LightGBM. Firstly, by introducing the Wasserstein distance as the loss function into the Conditional Generative Adversarial Network (CGAN), the Wasserstein Conditional Generative Adversarial Network (WCGAN) has been proposed to generate high-quality data, addressing the issue of insufficient information caused by small samples. Furthermore, by incorporating material and compaction parameters as inputs, a high-accuracy permeability coefficient prediction model was developed using LightGBM with the Huber loss function and the Osprey Optimization Algorithm (OOA). Finally, the SHapley Additive exPlanations (SHAP) method was introduced into OOA-HL-LightGBM to analyze the specific roles of different features within the dataset to enhance the credibility of the prediction results. The proposed method was applied to a large-scale high-core rockfill dam in Southwest China to thoroughly verify its effectiveness and superiority.

Key words: Permeability Coefficient Prediction; LightGBM; Wasserstein Conditional Generative Adversarial Nets; SHAP

Unedited

1 Introduction

The permeability coefficient is a crucial indicator for assessing the compaction quality of earthworks and is directly related to the structural stability and long-term operational safety of a project (Lin et al., 2023). During the compaction process, the traditional permeability coefficient testing method relies mainly on on-site testing at construction sites. However, this method has the following limitations: (1) the number of randomly selected points is often small, and the limited number of local permeability coefficients that reflect the permeability characteristics of the entire work area are unreliable; (2) the lengthy sampling and testing procedures are time-consuming, which may affect project progress; and (3) the on-site tests are carried out mainly at the end of the construction period, which makes it challenging to perform timely remedial action based on substandard test results (Liu et al., 2012; Lv et al., 2017).

Compared with traditional measurement methods, establishing a prediction model for the permeability coefficient and studying the nonlinear mapping relationships between the model and relevant parameters enables a real-time assessment of the permeability coefficient. With the development of artificial intelligence, machine learning algorithms have been widely used to predict permeability coefficients, including various neural networks (Liu et al., 2020; Wrzesinski and Markiewicz, 2022; Kim and Song, 2023), support vector machines (Bagheri and Rezaei, 2019; Seyyedattar et al., 2022), random forests (Zhao et al., 2023), and XGBoost (Nero et al., 2023). Overall, the use of machine learning techniques to estab-

✉ Hongling YU, yuhongling@cau.edu.cn
Chengyu YU, https://orcid.org/0009-0009-9126-7344
Hongling YU, https://orcid.org/0009-0009-1478-7338

Received Apr. 11, 2025; Revision accepted July 11, 2025; Crosschecked

© Zhejiang University Press 2025

lish models for predicting the permeability coefficient has matured. However, current research still faces the problems of small samples and the lack of explainability of the models.

Prediction problems with small samples are common in engineering applications. The nature of the small-sample prediction problem is the lack of information, which results in samples that do not adequately reflect the mapping relationship between the sample and label spaces (Fu et al., 2019). Data augmentation is a solution to the problem of insufficient sample data. Common data augmentation methods include Monte Carlo (Tsaregorodtsev and Belagiannis, 2023) based on sampling, Copula (Njock et al., 2025), Gaussian mixture model (Jiang et al., 2023) based on probability modeling, and variational autoencoder (Islam et al., 2021) based on generative models. However, these approaches have certain limitations in complex nonlinear dependencies among high-dimensional features. Generative adversarial networks (GANs), through the adversarial training mechanism and nonlinear modeling capacity of deep neural networks, can learn complex data distributions from the latent space and generate high-quality samples, making them widely used in small-sample problems (Chen et al., 2021; Wen et al., 2023; Mashhadi et al., 2024). However, the original GAN relies mainly on random noise to generate data (Han et al., 2024), making it uncontrollable and unable to produce results under the given conditions. The conditional generative adversarial network (CGAN) (Mirza and Osindero, 2014) is a variant of GANs, which constrains the model by adding conditional information ε in the generator and the discriminator thereby guiding the generator to produce data in a specified form. Moreover, the GAN training process is unstable and often suffers from convergence difficulties, pattern collapse, and gradient vanishing problems (Jabbar et al., 2021). To address these issues, Arjovsky et al. (2017) proposed the Wasserstein generative adversarial network (WGAN), which uses the Wasserstein distance as a measure for distribution discrepancies, replacing the Jensen-Shannon divergence in the conventional GAN. The superiority of the Wasserstein distance lies in its ability to reflect the dissimilarity between two distributions, even when the generated data distribution differs significantly from the original data distribution. This enables the effective alleviation of instability during training and the production of high-quality data without complicating the network model structure (Sang and Xu, 2022). The introduction of the Wasserstein distance as the loss function for the CGAN simultaneously enhances data controllability and improves training stability by enabling the generation of high-quality samples that approximate the real data distribution, thereby expanding the dataset and effectively mitigating the small sample problems.

Fully capitalizing on data information is key to improving prediction accuracy. Although neural networks have shown promising performance in engineering prediction tasks in recent years (Liu et al., 2025; Zhang et al., 2025), their prediction capability remains limited under small sample conditions. In contrast, ensemble learning algorithms, by integrating multiple weak learners, can mitigate overfitting more effectively and extract key information from the data more comprehensively, thereby obtaining better predictive performance (Zhou, 2023). Among ensemble learning algorithms, LightGBM demonstrates exceptional capabilities. It not only inherits the advantages of traditional ensemble learning methods but also shows excellent flexibility in handling various complex datasets (Ke et al., 2017). In recent years, LightGBM has been widely applied in many fields, including energy (Guo et al., 2023a), flood control (Ding et al., 2023), construction (Meng et al., 2024), and finance (Wang et al., 2022).

However, despite the excellent performance of LightGBM in handling complex information, outliers that may arise can still significantly impact prediction accuracy. In traditional regression losses, although the mean square error (MSE) provides good convergence, it assigns higher weights to outliers, affecting the overall performance of the model (Elsebach, 1994). While the mean absolute error (MAE) performs well in handling outliers, it is non-smooth at the global minimum, potentially hindering model convergence (Ahmed, 2023). The Huber loss function effectively combines the advantages of MSE and MAE. It mitigates sensitivity to outliers while maintaining differentiability, thereby enhancing model stability (You and Lu, 2021; Xing and Zhang, 2022; Yang et al., 2022). Inspired by this, in this study we introduced the Huber loss function into LightGBM to enhance its capability to handle outliers and improve its predic-

Unedited

tive performance.

Moreover, the choice of hyperparameters significantly affects the LightGBM's prediction results, making the selection of hyperparameters crucial. The osprey optimization algorithm (OOA), a new swarm intelligence optimization algorithm proposed in 2023, offers outstanding global search capability and fast convergence (Dehghani and Trojovský, 2023). Therefore, applying OOA to optimize the hyperparameters of LightGBM is an effective approach to improve the model's predictive performance.

Although the LightGBM algorithm can achieve high-precision model evaluation, its explainability remains limited, hindering a deeper understanding of the model decision (Sun et al., 2023a). Its explainability is manifested mainly in ranking feature importance. However, this approach is contingent upon how the model learns feature weights during training, constrained directly by the model's structure and learning capabilities. In recent years, scholars have used different approaches to study the explainability of machine learning, such as Partial Dependence Plot (PDP) (Friedman, 2001), Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), and SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017). The SHAP algorithm has garnered significant attention due to its powerful visualization capabilities and its ability to provide explanations of the model from both global and local perspectives. Many scholars have combined SHAP with various ensemble learning algorithms to improve the transparency and explainability of model decisions (Salehi et al., 2023; Sun et al., 2023b; Cakiroglu et al., 2024). The SHAP method shows a wide range of applicability. Applying it to explain LightGBM models by analyzing the impact of input features on predictions can significantly enhance model transparency.

In summary, the main contributions of this study are as follows:

(1) We propose a data augmentation technique based on WCGAN, guiding sample generation through CGAN and introducing Wasserstein distance as the loss function for CGAN, which addresses small-sample problems to a certain extent.

(2) The Huber loss is used as the loss function for LightGBM to mitigate the impact of outliers on prediction accuracy, with the OOA used to optimize the hyperparameters of the LightGBM ensemble learning algorithm. A permeability coefficient prediction model based on OOA-HL-LightGBM has been constructed, achieving high-precision prediction of permeability coefficients.

(3) The explainable OOA-HL-LightGBM permeability prediction model was applied to a large-scale high-core rockfill dam in Southwest China. SHAP was used to identify the most critical features affecting prediction results and to analyze the specific roles of different features within the samples, improving the credibility of the predictions. This application thoroughly verified the effectiveness and superiority of the proposed method.

The remainder of this paper is organized as follows: Section 2 describes the framework of this study. Section 3 describes the fundamentals of the explainable OOA-HL-LightGBM model with data augmentation. Section 4 applies the proposed method to engineering examples. Section 5 demonstrates the effectiveness and superiority of the technique through comparative analysis. Section 6 concludes this study and discusses future work.

Unedited

2 Research framework

Fig. 1 shows the research framework, which includes the following steps: (1) Acquiring relevant data from the database, (2) Developing a new method for predicting the permeability coefficients of earth-rock material based on the WCGAN data augmentation technique and explainable OOA-HL-LightGBM, and (3) The proposed method was applied to a large-scale high-core rockfill dam, followed by analysis and discussion.

3 Methodology

3.1 Data augmentation technology based on WCGAN

GAN is a type of deep neural network model trained in an adversarial manner, as proposed by Goodfellow et al. (2014), based on the concepts of games and adversarial training. A GAN consists of two functional networks: a generator and a discriminator. During practical training, the generator aims to produce fake samples that resemble an actual distribution, whereas the discriminator strives to enhance

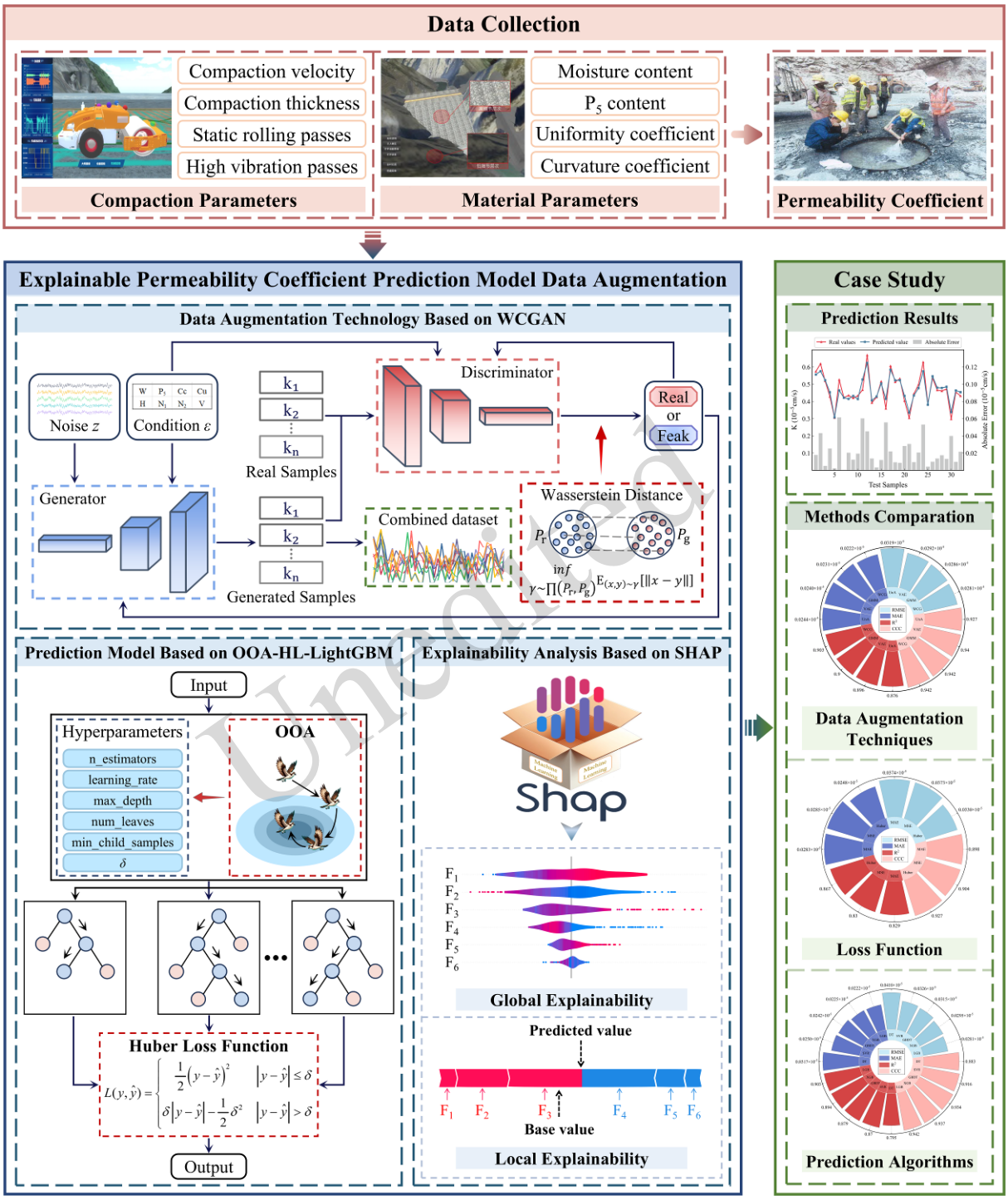


Fig. 1 Research Framework

Unedited

its accuracy in distinguishing between real and fake samples. This process continues to alternate, and make fake samples. The generator and discriminator continue improving their performance in the confrontation and finally generate data that approximate the truthful distribution. The training process can be expressed as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_r}[\log(D(x))] - E_{z\sim p_z}[\log(1-D(G(z)))], \quad (1)$$

where $V(D,G)$ denotes the value function of the

game, $E(\cdot)$ is the expectation operator, x is the samples from the real data distribution p_r, and z is noise sampled from the latent distribution p_z, $G(\cdot)$ is the generator function, which generates new data samples by taking a random noise vector from the latent space as input, and $D(\cdot)$ is the discriminator function, which takes a data sample as input and outputs the probability that the sample belongs to the real data distribution. $\min_G \max_D$ reflects the adversarial training process, in which the generator and discriminator gradually improve their performance through competition, ultimately enabling the generation of high-quality data samples.

CGAN is a variant based on GANs, which incorporates conditional information ε into both the generator and the discriminator, allowing the model to generate specified forms of data more quickly (Mirza and Osindero, 2014). The objective function is expressed as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_r}[\log(D(x\,|\,\varepsilon))] - E_{z\sim p_z}[\log(1-D(G(z\,|\,\varepsilon)))], \tag{2}$$

WGAN transforms this into an optimization problem that minimizes the Wasserstein distance by modifying the loss function of the GAN (Arjovsky, et al., 2017). During training, the generator attempts to minimize this function to generate data that are as similar as possible to the real data. The Wasserstein distance was calculated as follows:

$$W\left(P_r, P_g\right) = \inf_{\gamma\sim\prod(P_r,P_g)} E_{(x,y)\sim\gamma}\left[\|x-y\|\right], \tag{3}$$

where $\prod(P_r,P_g)$ represents the set of all possible joint distributions combining the real distribution P_r and the generated distribution P_g, and $\|x-y\|$ represents the distance between the real data and the generated data.

To facilitate the introduction of Wasserstein distance into GANs, a 1-Lipschitz constraint is added to the distance calculation, aiming to ensure that the output curve remains as smooth as possible without tending toward infinity or infinitesimal. Additionally, to satisfy Lipschitz continuity in the training network, weight clipping is applied, restricting the network parameters to a predefined range after each update. After introducing the Wasserstein distance and Lipschitz constraints, the objective function of WGAN is as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_r}[D(x)] - E_{z\sim p_z}[D(G(z))], \tag{4}$$

In this study, we combined the advantages of CGAN and WGAN to propose an improved GAN-based structure, namely WCGAN. In WCGAN, conditional information ε is incorporated into both the generator and discriminator, while the Wasserstein distance is used as the loss function. The objective function is as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_r}[D(x\,|\,\varepsilon)] - E_{z\sim p_z}[D(G(z\,|\,\varepsilon))], \tag{5}$$

Based on this structure, WCGAN is used to perform data augmentation, and its working principle is illustrated in Fig. 2. The input features, including material and compaction parameters, are first combined to form the conditional information ε, which is used to guide the generation process. During training, the generator takes a concatenation of ε and a random noise z as input and outputs the corresponding permeability coefficient. The discriminator receives either a real or a generated permeability coefficient, along with the same conditional information ε, and

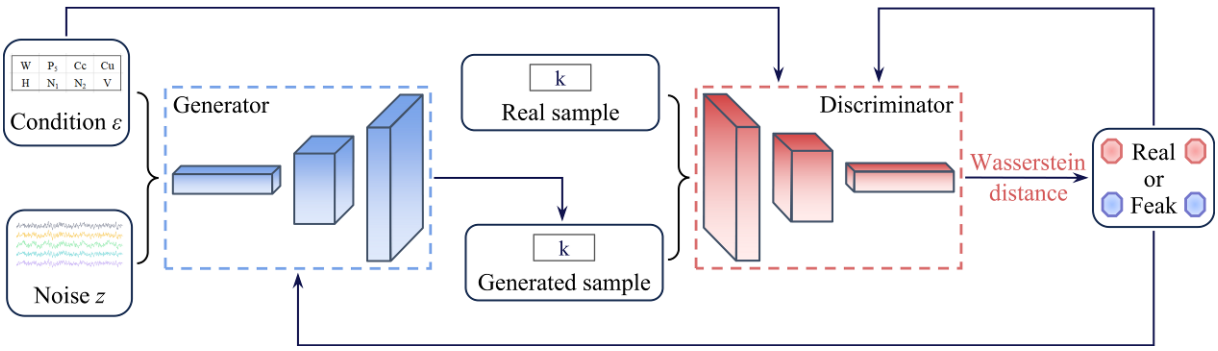


Fig. 2 Schematic diagram of the working principle of WCGAN.

Unedited

determines whether the input is from the real data distribution. After training, new input features are generated by randomly sampling the original features with replacement and introducing small perturbations. These features are then combined to construct conditional information ε, which is concatenated with random noise z and input to the generator to produce the corresponding permeability coefficients.

3.2 Permeability coefficient prediction model based on OOA-HL-LightGBM

3.2.1 Huber Loss Function

Huber loss is a loss function that effectively mitigates the impact of outliers on the computational results (Huber, 1964). The mathematical expressions are as follows:

$$L(y,\hat{y})=\begin{cases}\frac{1}{2}\left(y-\hat{y}\right)^2 & \left|y-\hat{y}\right|\le\delta\\ \delta\left|y-\hat{y}\right|-\frac{1}{2}\delta^2 & \left|y-\hat{y}\right|>\delta\end{cases}, \tag{6}$$

where δ is a tunable parameter, and $y-\hat{y}$ represents the error between the actual value and the predicted value. For small errors ($\left|y-\hat{y}\right|\le\delta$), the Huber loss approximates MSE loss, with the gradient decreasing as the error decreases, facilitating the model's ability to converge more precisely to the optimal point. For large errors ($\left|y-\hat{y}\right|>\delta$), the loss function transitions to the MAE loss form, with a gradient approximately equal to δ, ensuring that the model updates parameters at a faster rate.

3.2.2 LightGBM ensemble learning algorithm

LightGBM is an ensemble learning method optimized for the Gradient Boosting Decision Tree (GBDT) model (Ke et al., 2017). It introduces Gradient-based One-Sided Sampling (GOSS) to minimize the required number of samples while ensuring prediction accuracy, thereby accelerating training and reducing memory usage. Meanwhile, the model uses Exclusive Feature Bundling (EFB), effectively reducing the number of features and lowering the dimensionality, enhancing its capability to handle high-dimensional data (Guo et al., 2023b). Furthermore, the use of histogram algorithms significantly reduces the time complexity and doubles the computational speed (Wang et al., 2022). LightGBM also uses a leaf-wise strategy with depth limitations to reduce errors and achieve higher accuracy (Wang and Wang, 2020).

The application of these four innovative technologies endows the LightGBM with unique advantages for handling high-dimensional feature data (Essa et al., 2023). The LightGBM constructs predictive models by integrating multiple decision trees. Specifically, for the i-th sample x_i, the predicted value $\widehat{y_i}$ at the t-th iteration is represented as the cumulative sum of the predictions of the first t trees, which is mathematically expressed as:

$$\widehat{y_i}^{(t)}=\sum_{k=1}^{t}f_k\left(x_i\right), \tag{7}$$

where $f_k(x_i)$ denotes the prediction result of the k-th regression decision tree for the i-th sample. Moreover, the loss function $L^{(t)}$ in each iteration considers the difference between the actual labels and current model predictions by incorporating regularization terms to prevent overfitting. After using the Huber loss, the expression of the loss function for the LightGBM is as follows:

$$L^{(t)}=\sum_{i=1}^{n}l\left(y_i,\widehat{y_i}^{(t-1)}+f_{t(x_i)}\right)+\Omega\left(f_t\right) \tag{8}$$

$$\Omega\left(f_t\right)=\gamma T+\frac{1}{2}\lambda\left\|\boldsymbol{\omega}\right\|^2 \tag{9}$$

where $l(\cdot)$ represents the loss function, used to measure the loss between the actual label y_i and the predicted value of the model after incorporating the t-th tree. $\Omega(f_t)$ represents the regularization term for the t-th tree, while γ and λ are the regularization coefficients for the number of leaf nodes and their respective leaf node weights, respectively.

3.2.3 Osprey optimization algorithm (OOA)

OOA is a metaheuristic algorithm that optimizes hyperparameters by simulating the predatory behavior of ospreys in nature (Dehghani and Trojovský, 2023). Ospreys are birds that feed on various fish, and their hunting behavior is divided into two phases: locating the fish and bringing the fish to a safe place. In OOA, the position of each osprey is considered a candidate solution to the problem. As the osprey changes its position during fishing, if the new position corresponds to a better value of the objective function, it replaces the os-

Unedited

prey's previous position. This process is mathematically expressed as:

$$\boldsymbol{X}_i \begin{cases} \boldsymbol{X}_i^{P_n} & F_i^{P_n} < F_i \\ \boldsymbol{X}_i & else \end{cases}, \ n = 1,\ 2 \tag{10}$$

where $\boldsymbol{X}_i$ is the position of the i-th osprey, F_i is the corresponding fitness function, and P_n denotes the fishing phases of the osprey.

In the first phase, the osprey randomly discovers the position of a fish and attacks it. This process causes significant changes in the osprey's position in the search space, constituting the global exploration phase of OOA. The movement of the osprey towards the school of fish and the updating of its position is simulated using the following formula:

$$\begin{gathered} X_{i,j}^{P_1} = X_{i,j} + r_{i,j} \cdot \left(SF_{i,j} - I_{i,j} \cdot x_{i,j} \right), \\ i = 1,\ 2,\ \cdots,\ N.\ j = 1,\ 2,\ \cdots,\ m. \end{gathered} \tag{11}$$

where $X_{i,j}^{P_1}$, j represents the new position in the j-th dimension of the i-th osprey in the first phase, SF_i is a random number between [0, 1], and $I_{i,j}$ is a random number from set {0, 1}.

After catching a fish, the osprey takes it to a safe location to eat. This process results in minor changes in the osprey's position in the search space, representing the local development phase of OOA. This change helps to avoid getting trapped in local optima and enhances the OOA's capability in local search development. The osprey's position change process in this phase is illustrated by the following equation:

$$\begin{gathered} X_{i,j}^{P_2} = X_{i,j} + \frac{lb_j + r \cdot \left(ub_j - lb_j \right)}{t}, \\ i = 1,\ 2,\ \cdots,\ N.\ j = 1,\ 2,\ \cdots,\ m.\ t = 1,\ 2,\ \cdots,\ T. \end{gathered} \tag{12}$$

where $X_{i,j}^{P_2}$ represent the new positions in the j-th dimension of the i-th osprey in the second phase. The variable r is a random number between [0, 1], and ub_j and lb_j are the upper and lower bounds of the optimization. The variable t is the current iteration number, and T is the maximum number of iterations, respectively.

3.3 Additive feature explanation method based on SHAP

SHAP is a model-agnostic additive feature explanation method proposed by Lundberg and Lee (2017), based on the concepts of cooperative game theory. Its essence lies in the Shapley values introduced by Shapley (1952), which enable the quantitative assessment of each feature's contribution to the prediction outcome, thus effectively explaining complex machine learning models. In predictive models, the formula for calculating the Shapley value of the i-th input feature is as follows:

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!\left(|M| - |S| - 1\right)!}{|M|!} \left[f_{S \cup \{i\}}\left(\boldsymbol{x}_{S \cup \{i\}} \right) - f_S\left(\boldsymbol{x}_S \right) \right] \tag{13}$$

where M is the set that includes all input features, $|M|$ denotes the number of feature dimensions in M, S represents all subsets of features excluding features i, $|S|!$ is the factorial of the number of features in S, $\boldsymbol{x}_S$ denotes the input feature values in S, $f_{S \cup \{i\}}$ is a model trained including feature i, and f_S is another model trained excluding feature i.

Based on Shapley values, SHAP approximates complex models by combining multiple linear models, essentially defining the output model as the sum of contributions from input variables (Oliveira et al., 2024). Therefore, the explanation of the predictive model by SHAP can be represented by the following equation:

$$g(\boldsymbol{z}') = \phi_0 + \sum_{i=1}^{m} \phi_i z_i', \tag{14}$$

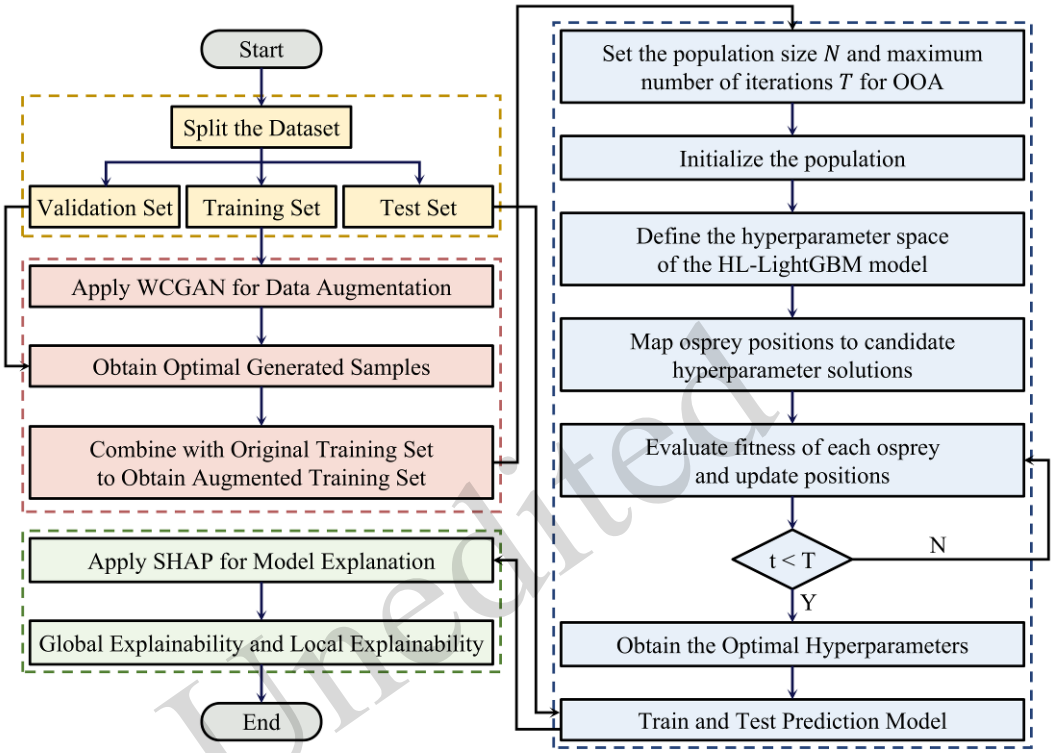
where ϕ_0 is the baseline output of the entire predictive model, ϕ_i is the Shapley value of the i-th feature, and z_i' indicates whether the i-th feature is selected, with z_i' equaling 1 when the feature is selected, and 0 otherwise.

3.4 Prediction process of the permeability coefficient under small sample conditions using the improved GAN and explainable ensemble learning

A flowchart of the proposed method is illustrated in Fig. 3. Firstly, the original dataset is split into a training set, a validation set, and a test set. Then, WCGAN is applied to augment the training set. The generated samples are evaluated using the validation set, and the optimal ones are selected and combined with the original training data to form an

Unedited

augmented training set. Subsequently, OOA is used to optimize the hyperparameters of the HL-LightGBM model. After obtaining the optimal hyperparameter combination, the permeability coefficient prediction model is trained using the augmented training set, and its performance is evaluated on the test set. Finally, SHAP is applied to perform both global and local explanations of the model.



Unedited

Fig. 3 Flowchart of the proposed method for permeability coefficient prediction.

4 Case Study

A case study was conducted on a high-core rockfill dam in Southwest China, which focuses mainly on power generation along with flood control, and has multi-year regulation capabilities. Given the critical role of the core wall area in seepage prevention, we analyzed the permeability coefficient of earth-rock materials during the compaction process.

4.1 Data Collection

Compaction construction information comes from numerous sources and is characterized by different temporal granularities and spatial locations, making it a typical case of multi-source heterogeneous information. To construct the original input and output space for a permeability coefficient predictive model, it is necessary to perform spatiotemporal pairing of this multi-source heterogeneous compaction construction information.

As illustrated in Fig. 4, the compaction parameters were obtained from a real-time monitoring system developed by Zhong et al. (2011). This system, based on the Global Navigation Satellite System (GNSS) positioning module and the vibration force monitoring module, obtains compaction velocity (V), compaction thickness (H), static rolling passes (N_1) and high-vibration passes (N_2)-through data reading and processing by the database server and application server. For the material parameters, the moisture content (W) was obtained from moisture content tests, the P_5 content (P_5), uniformity coefficient (C_u), and curvature coefficient (C_c) were obtained from particle analysis tests, and the permeability coefficient was obtained from spot tests. Further data collection and compilation in the database were conducted using a PDA (personal digital assistant) real-time information collection system (Liu et al., 2015).

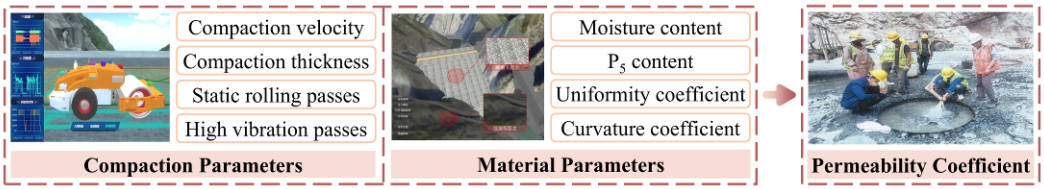


Fig. 4 Data collection of compaction parameters, material parameters and permeability coefficient.

Considering that the compaction construction information collected in its original state lacks direct interconnections, we transformed multisource heterogeneous compaction construction information to the same spatiotemporal baseline based on spatiotemporal correlations. This transformation achieves the spatiotemporal pairing of the input and output data, resulting in an original dataset. From this dataset, 207 sets of data were selected for predictive analysis of the permeability coefficient. These sets were split in a 7:1.5:1.5 ratio into 144 training samples, 31 validation samples, and 32 test samples.

4.2 Analysis of Data Augmentation Technology

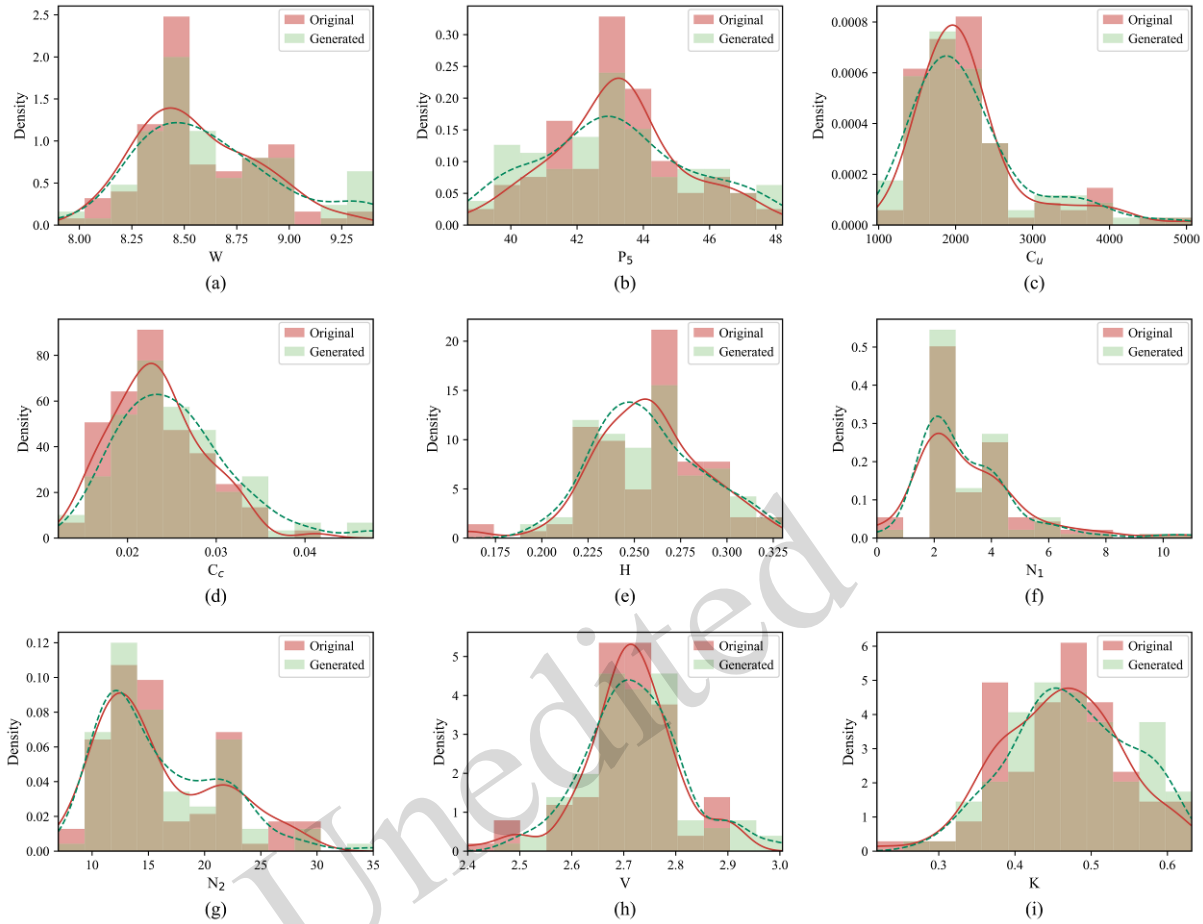
Using the WCGAN method proposed in this study, 25, 50, 100, 150, and 175 samples were generated from the original training set. Subsequently, these samples were combined with the original training data to form a new training set. The previously defined validation set was used for testing. A prediction model constructed using the LightGBM algorithm with the Huber loss function was applied to predict the permeability coefficient. The prediction results for the training data containing different generated samples are shown in Table 1. The four evaluation metrics—root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and concordance correlation coefficient (CCC) (Lin, 1989)—showed good consistency. Compared to the prediction model without added generated samples, the model with added generated samples had lower RMSE and MAE, and higher R^2 and CCC values, indicating that the model with added generated samples has higher prediction accuracy. Moreover, among the models with added generated samples, as the number of generated samples increased, the predictive performance of the model initially improved and then diminished. This indicates that although using WCGAN to generate samples for data augmentation can improve the model's generalizability, adding more generated samples does not necessarily improve the model's predictive capability. This is because generating too many samples may lead to overfitting on the training set and reduced generalizability on real data.

Table 1. Comparison of results for different numbers of generated samples.

Number of samples	144	144+25	144+50	144+100	144+150	144+175
RMSE	0.0444	0.0410	0.0403	0.0374	0.0382	0.0397
MAE	0.0352	0.0306	0.0323	0.0276	0.0311	0.0299
R^2	0.813	0.840	0.845	0.867	0.861	0.850
CCC	0.896	0.907	0.913	0.921	0.920	0.914

To further verify the statistical consistency of the distribution characteristics of the generated samples, Fig. 5 presents a comparison between the original samples and the generated samples under the optimal generation size in terms of histograms and Gaussian kernel density estimation curves across each feature dimension. The density curves of the original and generated data show similar distributions, with similar trends of variation across different value ranges. Meanwhile, the frequency distributions shown in the histograms also demonstrate a high degree of consistency. These results indicate that the WCGAN is capable of effectively learning the statistical characteristics of the original data and generating samples with a similar distribution.

Unedited



Unedited

Fig. 5 Data distribution of original data and generated data (the brown color indicates overlapping regions).

4.3 Analysis of Permeation Coefficient Prediction Results

Combining the best-performing 100 generated samples with the original training data formed an augmented training set containing 244 samples. To ensure consistency, the 32 samples designated as the test set in the earlier split were used for model evaluation. The model constructed using the LightGBM algorithm with the Huber loss function was applied to predict the permeability coefficient.

The hyperparameters of the HL-LightGBM algorithm were adaptively optimized using OOA. In the OOA algorithm, the osprey population size was 40, and the maximum iteration number was set to 200. R^2 was used as the fitness function to maximize the predictive accuracy of the model. To ensure robustness and prevent data leakage, five-fold cross-validation was used during the OOA-based hyperparameter optimization of the HL-LightGBM model. The average R^2 score served as the evaluation metric for assessing the performance of each hyperparameter combination. The optimization ranges and optimal results for various parameters are shown in Table 2.

Table 2. Optimized results for OOA-HL-LightGBM.

Parameters	Optimization scopes	Optimal results
n_estimators	[1, 500]	130
learning_rate	[0.01, 1]	0.28
max_depth	[2, 100]	11
num_leaves	[2, 100]	38
min_child_samples	[2, 100]	27
δ	[0.01, 5]	0.03

After determining the optimal combination of hyperparameters, the model was trained using an augmented training set containing 244 samples. Subsequently, the predictive performance of the

model was validated with the test set. The prediction results of the test set (Fig. 6) indicate that the predicted values of the permeability coefficient closely match the actual values. This shows that the permeability coefficient model using OOA-HL-LightGBM ensemble learning has high prediction accuracy. Further calculations showed that the RMSE, MAE, R^2, and CCC were 0.0281×10^{-5}cm/s, 0.0222×10^{-5}cm/s, 0.903, and 0.942, respectively. These metrics collectively indicate that the model has high prediction accuracy and strong generalization capability.

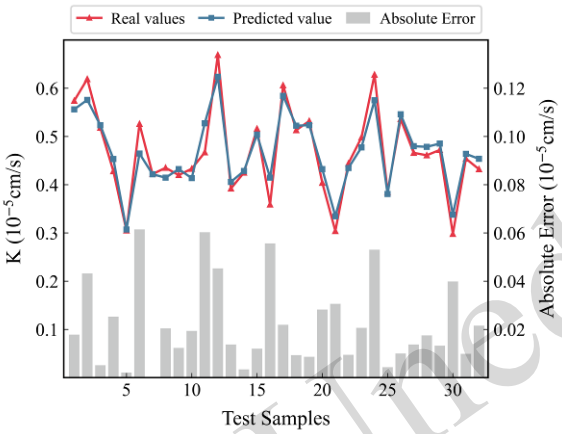


Fig. 6 Permeability coefficient prediction results.

4.4 Explainability Analysis of the Permeability Coefficient Prediction Model

4.4.1 Global Explainability Analysis

The SHAP method calculates the Shapley value of each feature using its average absolute value as an indicator of feature importance. The larger this indicator value, the higher the importance of the feature. The SHAP feature importance ranking of the OOA-HL-LightGBM-based permeability coefficient prediction model is shown in Fig. 7. P_5 had the most significant impact on the permeability coefficient prediction. N_2 and H played essential roles in the predictions of the model and C_c also influenced the prediction to some extent. N_1, W, C_u, and V had a weaker impact than C_c.

Additionally, the Shapley value for each feature across all samples was plotted as a scatter plot, with a color gradient indicating feature magnitudes (red for high, blue for low). Each point represents a Shapley value for a particular sample feature. The y-axis lists features by importance, while the x-axis shows Shapley values, where positive values indicate a positive contribution and negative values indicate a negative impact. With an increase in P_5, the corresponding Shapley value was less than zero and continued to decrease (Fig. 8), indicating that higher P_5 harms the prediction of the permeability coefficient. Similar trends were observed for C_u and V. In contrast, an opposite trend was observed for N_2. As the feature value increased, its corresponding Shapley value became greater than zero and continued to increase, indicating that larger feature values resulted in higher predictions of the permeability coefficient. Similar patterns were observed for H, C_c, N_1, and W.

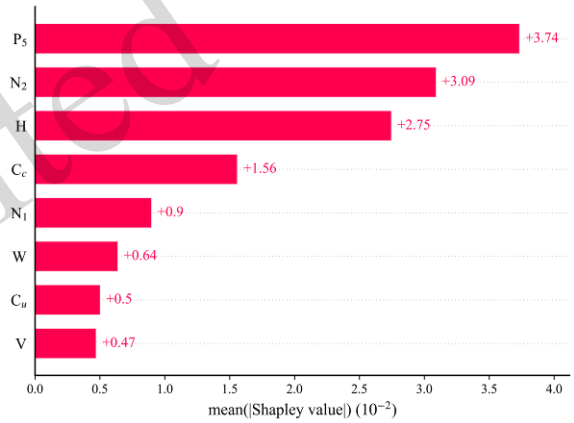


Fig. 7 Feature importance ranking based on SHAP.

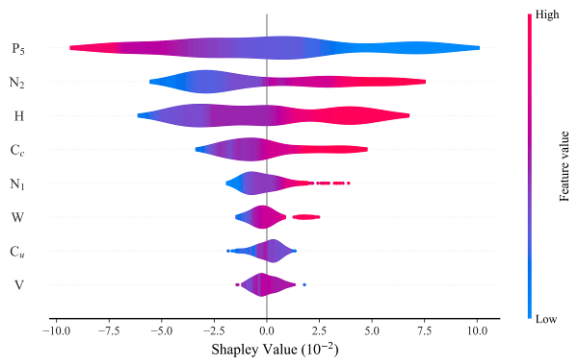


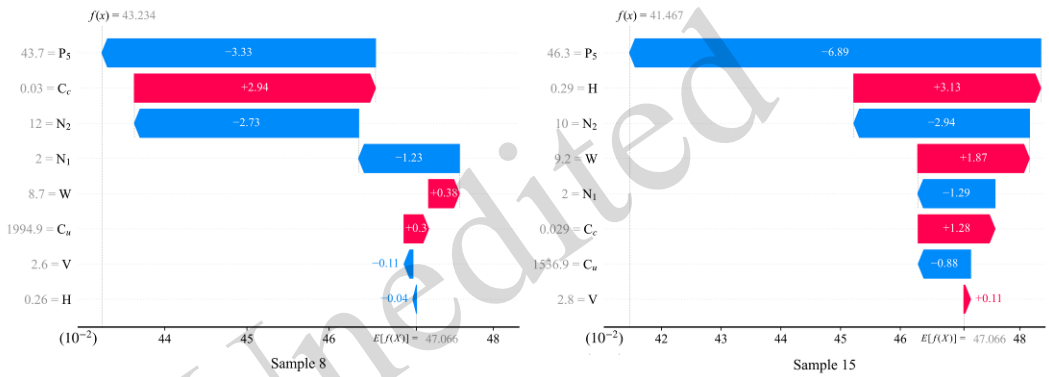
Fig. 8 SHAP feature importance violin plot.

4.4.2 Local Explainability Analysis

The SHAP attribution analysis method not only provides a comprehensive analysis of the importance of sample features globally and identifies key factors affecting prediction outcomes but also enables the analysis of the effects of different features in individual samples. Using Samples 8 and 15 as exam-

Unedited

ples, the effects of their features is illustrated in Fig. 9. Here, the x-axis represents the magnitude of the predicted values, and the y-axis represents the input feature values arranged by their importance. Red indicates a positive impact on the prediction of the permeability coefficient, whereas blue indicates a negative effect. For Sample 8, P_5 is the main feature influencing the prediction output, and Cc and N_2 also have a significant impact. P_5, N_2, N_1, V, and H hinder the prediction of the permeability coefficient, while C_c, W, and C_u have positive effects. Specifically, a P_5 value of 43.7 decreased the permeability coefficient from the baseline by 3.33×10^{-7} cm/s, and a C_c value of 0.03 increased the predicted value from the baseline by 2.94×10−7 cm/s. The baseline permeability coefficient, representing the model's average prediction across the training dataset, was 47.0666×10^{-7} cm/s. Accounting for the combined contribution of all features, the final predicted permeability coefficient for Sample 8 was 43.234×10^{-7} cm/s. In addition, the comparison between Sample 8 and Sample 15 reveals that the same input feature may exert different effects across samples, indicating that the explanations and analyses of input features should be context-specific rather than overly absolute.



Unedited

Fig. 9 Single sample prediction diagram.

SHAP also enables the visualization of the interactive effects of two different features on the Shapley value through feature dependence plots. The horizontal axis represents the value of the input feature, the left vertical axis shows the Shapley value of the input feature, and the right vertical axis represents the value of the other feature. The redder the color, the larger the feature value, and the bluer the color, the smaller the feature value.

The interactive effects of H and P_5 are shown in Fig. 10. When H was greater than 0.26, most of the Shapley values were greater than 0, indicating that an

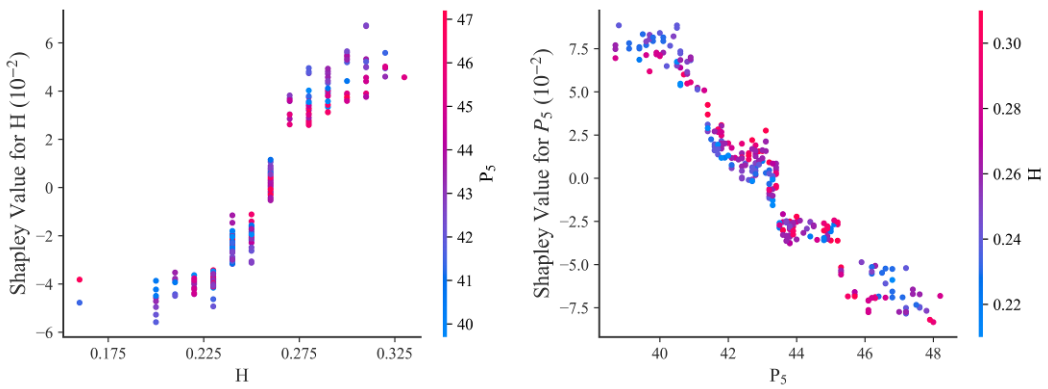


Fig. 10. The interaction between *H* and *P*$_5$.

increase in H positively impacted the prediction of the permeability coefficient in this range. Conversely, when H was less than 0.26 most Shapley values were less than 0, suggesting that a decrease in

H negatively influenced the prediction of the permeability coefficient. Similarly, when P_5 exceeded 43.5, most of its Shapley values were less than 0, negatively affecting the prediction of the permeability coefficient, and when P_5 was below 43.5, most of its Shapley values were greater than 0, positively impacting the prediction. Through a comprehensive analysis of the interaction between H and P_5, we observed that when H was greater than 0.26, a lower P_5 increased the positive contribution of H to the predicted permeability coefficient. When H was less than 0.26, a lower P_5 increased its negative contribution. Furthermore, when P_5 exceeded 43.5, a higher H enhanced its negative effect. When P_5 ranged from 41 to 43.5, a higher H enhanced its positive effect, while when P_5 was below 41, a lower H enhanced its positive effect.

5 Discussion

5.1 Comparative Analysis of Data Augmentation Technology

To validate the effectiveness of the proposed WCGAN-based data augmentation method in addressing small sample prediction issues in practical engineering, we compared it with alternative data augmentation techniques based on Gaussian Mixture Models (GMM) combined with K-Nearest Neighbors (KNN) regression, as well as Variational Autoencoders (VAE). Following the strategy described in Section 4.2, multiple sets of samples were generated using the GMM-KNN-based and VAE-based data augmentation methods. The optimal number of samples generated for each method was selected based on the prediction performance on the validation set. These samples were combined with the original training data to form the final training set, on which the model was trained. The prediction performance was evaluated using the test set. The prediction results of the samples generated by the three different methods, along with those from the original training data (UnA), are shown in Fig. 11.

Analysis showed that the three data augmentation methods demonstrated great consistency across the four evaluation metrics of RMSE, MAE, R^2, and CCC. The prediction model trained with the training set expanded by the WCGAN-based method significantly outperformed the other two methods across all evaluation metrics. Moreover, compared to the original samples, adding WCGAN-generated samples reduced the model's MAE by 9%, outperforming the reductions of 5.3% and 1.7% achieved with the GMM-KNN and VAE-based methods, respectively. This indicates that the proposed WCGAN has superior generative capabilities, effectively enhancing the model's predictive performance.

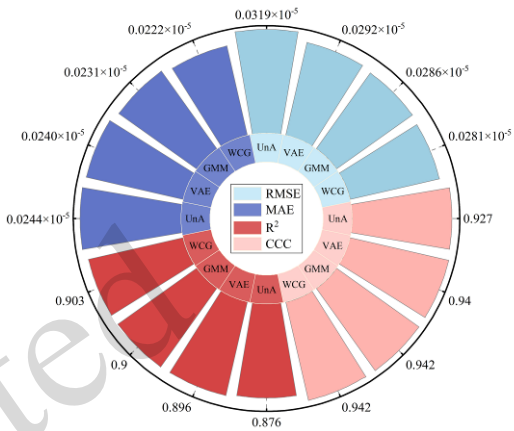


Fig. 11 Comparison of different data augmentation techniques.

5.2 Comparative Analysis of Prediction Model Performance

5.2.1 Comparative Analysis of Loss Function

To explore the predictive performance of models using the Huber loss function, the results of the LightGBM permeability prediction model using Huber loss as the loss function were compared with those of models using the MSE and MAE loss function. The results are shown in Fig. 12.

The analysis indicated that the predictive model using the Huber loss function significantly outperformed the other two methods in terms of RMSE, MAE, R^2, and CCC. The MAE of the model using Huber loss decreased by 13.0% and 12.4% compared to the models using MSE and MAE losses, respectively, demonstrating that the Huber loss function provides the model with superior predictive performance.

5.2.2 Comparative Analysis of Prediction Algorithm

To validate the predictive performance of the LightGBM ensemble learning model, we compared it with four other prediction models, including DT, SVR, GBDT, and XGBoost, all of which were optimized using the OOA-based hyperparameter tuning

Unedited

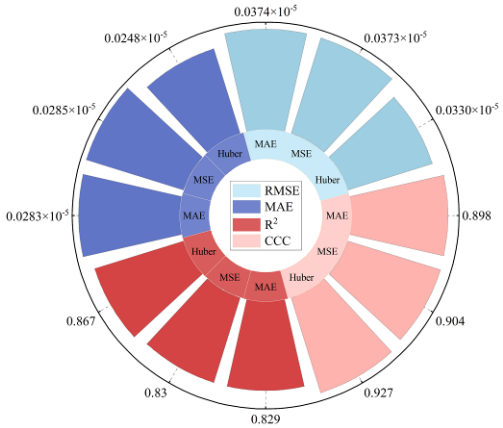


Fig. 12 Comparison of predictive results using different loss functions.

approach described in Section 4.3. As shown in Fig. 13, it is evident that compared with the other four algorithms, the LightGBM ensemble learning algorithm had the highest R^2 and CCC, but the lowest RMSE and MAE. Further calculations show that compared to DT, SVR, GBDT, and XGBoost, the LightGBM model achieved MAE reductions of 30%, 11.2%, 8.3%, and 1.3%, respectively, indicating that LightGBM has excellent prediction accuracy. Additionally, the prediction accuracies of DT and SVR were lower than those of GBDT and XGBoost. This is because DT and SVR are individual machine learning algorithms, while GBDT and XGBoost are ensemble learning algorithms with better predictive performance.

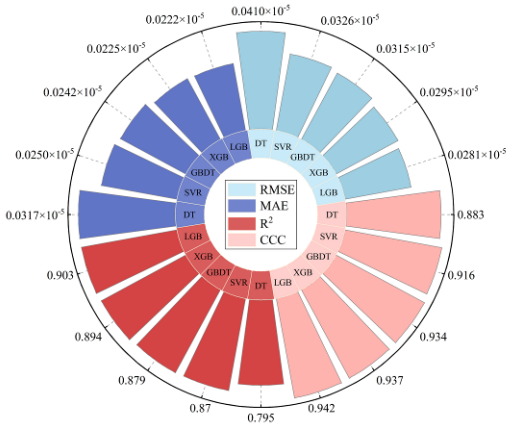


Fig. 13 Comparison of prediction results from different algorithms.

6 Conclusions

The permeability coefficient is a critical indicator for assessing the compaction quality of earthworks. In this paper, we propose a permeability coefficient prediction model that uses WCGAN data augmentation technology and explainable OOA-HL-LightGBM. The conclusions of this study are as follows:

(1) The proposed WCGAN guides sample generation through the CGAN mechanism and introduces Wasserstein distance as the loss function, which generates high-quality samples stably and addresses the issue of small sample data effectively. Compared with the methods based on GMM-KNN and VAE, the WCGAN-based method shows superior performance in improving the accuracy of predictive models.

(2) This method uses Huber loss as the loss function for the LightGBM ensemble learning algorithm and optimizes LightGBM's hyperparameters using OOA, thereby establishing a high-precision permeability coefficient prediction model. The results indicate that models using Huber loss have improved prediction accuracy compared to those using MSE and MAE losses. Furthermore, compared to traditional algorithms, LightGBM demonstrates advanced performance.

(3) This method integrates the SHAP attribution analysis method with the OOA-HL-LightGBM ensemble learning algorithm to enhance the explainability of the prediction model and increase the credibility of the prediction results.

In this paper, we propose a new method for predicting the permeability coefficient of earth-rock material. Currently, the method focuses mainly on the study of the permeability compaction characteristics. Future work will extend to simultaneously predicting the physical, mechanical, and permeability compaction characteristics. There are also plans to integrate this prediction model into intelligent compaction monitoring systems, to achieve real-time assessment of earthwork compaction quality. Moreover, considering the generalization potential of the methodological framework proposed in this paper, future work will explore its applicability to other geotechnical or civil engineering scenarios, such as subgrade compaction and embankment engineering, where the same type of compaction processes are involved. These engineering applications also require quantitative indicators to evaluate compaction quality and follow similar data acquisition procedures, making it possible to apply the proposed

Unedited

framework to new engineering prediction tasks by retraining the model with task-specific data.

Acknowledgments

This work was funded by the Youth Program of the National Natural Science Foundation of China (No. 52409181) and the National Natural Science Foundation of China (No. U23B20148).

Author contributions

Chengyu YU designed the research, analyzed and processed the corresponding data, validated the methodology, and wrote the first draft of the manuscript. Hongling YU contributed to the methodology and provided supervision and resources. Xiaofeng QU provided resources and participated in validation. Baoxi LIU provided supervision and resources. Liangsi XU, Xinyu LIU, and Xiangyu CHEN participated in validation. All authors contributed to the review and editing of the manuscript.

Conflict of interest

Chengyu YU, Hongling YU, Xiaofeng QU, Liangsi XU, Xinyu LIU and Xiangyu CHEN declare that they have no conflict of interest.

Unedited

References

Ahmed KJ, 2023. Batch-stochastic sub-gradient method for solving non-smooth convex loss function problems. International Journal of Mobile Network Communications Telematics, p.65-84. https://doi.org/10.5121/csit.2023.131806

Arjovsky M, Chintala S, Bottou L, 2017. Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning, JMLR.org, p.214–223.

Bagheri M, Rezaei H, 2019. Reservoir rock permeability prediction using svr based on radial basis function kernel. *Carbonates and Evaporites*, 34:699-707. https://doi.org/10.1007/s13146-019-00493-4

Cakiroglu C, Demir S, Ozdemir MH, et al., 2024. Data-driven interpretable ensemble learning methods for the prediction of wind turbine power incorporating shap analysis. *Expert Systems with Applications*, 237 https://doi.org/10.1016/j.eswa.2023.121464

Chen ZS, Hou KR, Zhu MY, et al., 2021. A virtual sample generation approach based on a modified conditional gan and centroidal voronoi tessellation sampling to cope with small sample size problems: Application to soft sensing for chemical process. *Applied Soft Computing*, 101 https://doi.org/10.1016/j.asoc.2020.107070

Dehghani M, Trojovský P, 2023. Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems. *Frontiers in Mechanical Engineering*, 8 https://doi.org/10.3389/fmech.2022.1126450

Ding F, Zhang WJ, Cao SH, et al., 2023. Optimization of water quality index models using machine learning approaches. *Water Research*, 243 https://doi.org/10.1016/j.watres.2023.120337

Elsebach R, 1994. Evaluation of forecasts in ar models with outliers. *Operations-Research-Spektrum*, 16(1):41-45. https://doi.org/10.1007/bf01719702

Essa E, Omar K, Alqahtani A, 2023. Fake news detection based on a hybrid bert and lightgbm models. *Complex & Intelligent Systems*, 9(6):6581-6592. https://doi.org/10.1007/s40747-023-01098-0

Friedman JH, 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189-1232. https://doi.org/10.1214/aos/1013203451

Fu XY, Luo H, Zhang GY, et al., 2019. A lazy support vector regression model for prediction problems with small sample size. *Neural Network World*, 29(1):33-44. https://doi.org/10.14311/nnw.2019.29.003

Goodfellow I, Pouget-Abadie J, Mirza M, et al., 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27

Guo JX, Yun SN, Meng Y, et al., 2023a. Prediction of heating and cooling loads based on light gradient boosting machine algorithms. *Building and Environment*, 236 https://doi.org/10.1016/j.buildenv.2023.110252

Guo MX, Guo Y, Peng YF, et al., 2023b. Fault diagnosis of bolt loosening based on lightgbm recognition of sound signal features. *Ieee Sensors Journal*, 23(19):22777-22787. https://doi.org/10.1109/jsen.2023.3303223

Han K, Yu Y, Lu T, 2024. Transfer learning and interpretable analysis-based quality assessment of synthetic optical coherence tomography images by cgan model for retinal diseases. *Processes*, 12(1) https://doi.org/10.3390/pr12010182

Huber PJ, 1964. Robust estimation of alocation parmneter. *Annals of Mathematical Statistics*, 35:73-101.

Islam Z, Abdel-Aty M, Cai Q, et al., 2021. Crash data augmentation using variational autoencoder. *Accident Analysis and Prevention*, 151 https://doi.org/10.1016/j.aap.2020.105950

Jabbar A, Li X, Omar B, 2021. A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, 54(8) https://doi.org/10.1145/3463475

Jiang XY, Yao L, Yang ZY, et al., 2023. Gaussian mixture model and double-weighted deep neural networks for data augmentation soft sensing. IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), IEEE, p.1914-1919. https://doi.org/10.1109/ddcls58216.2023.10166693

Ke GL, Meng Q, Finley T, et al., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3149–3157.

Kim MH, Song CM, 2023. Prediction of the soil permeability

coefficient of reservoirs using a deep neural network based on a dendrite concept. *Processes*, 11(3) https://doi.org/Processes

Lin LI, 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255-268. https://doi.org/10.2307/2532051

Lin WW, Wang JJ, Wang XL, et al., 2023. An enhanced multiobjective bacterial foraging algorithm for the compaction parameter optimization of earth-rock dams. *Construction and Building Materials*, 394 https://doi.org/10.1016/j.conbuildmat.2023.132178

Liu DH, Sun J, Zhong DH, et al., 2012. Compaction quality control of earth-rock dam construction using real-time field operation data. *Journal of Construction Engineering and Management*, 138(9):1085-1094. https://doi.org/10.1061/(asce)co.1943-7862.0000510

Liu HC, Zhang N, Yin ZY, 2025. Probabilistic stratigraphic modeling from sparse boreholes based on deep learning. *Geotechnique*, https://doi.org/10.1680/jgeot.24.00998

Liu XL, Li DL, Yang JH, et al., 2020. Automatic well test interpretation based on convolutional neural network for infinite reservoir. *Journal of Petroleum Science and Engineering*, 195 https://doi.org/10.1016/j.petrol.2020.107618

Liu YX, Zhong DH, Cui B, et al., 2015. Study on real-time construction quality monitoring of storehouse surfaces for rcc dams. *Automation in Construction*, 49:100-112. https://doi.org/10.1016/j.autcon.2014.10.003

Lundberg SM, Lee SI, 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30

Lv P, Wang XL, Liu Z, et al., 2017. Porosity-and reliability-based evaluation of concrete-face rock dam compaction quality. *Automation in Construction*, 81:196-209. https://doi.org/10.1016/j.autcon.2017.06.019

Mashhadi AH, Rashidi A, Markovic N, 2024. A gan-augmented cnn approach for automated roadside safety assessment of rural roadways. *Journal of Computing in Civil Engineering*, 38(2) https://doi.org/10.1061/jccee5.Cpeng-5406

Meng CC, Qu DY, Duan XC, 2024. Cost estimation of metro construction projects using interpretable machine learning. *Journal of Computing in Civil Engineering*, 38(6) https://doi.org/10.1061/jccee5.Cpeng-6018

Mirza M, Osindero S, 2014. Conditional generative adversarial nets. *arXiv e-prints*, arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

Nero C, Aning AA, Danuor SK, et al., 2023. Prediction of compressional sonic log in the western (tano) sedimentary basin of ghana, west africa using supervised machine learning algorithms. *Heliyon*, 9(9) https://doi.org/10.1016/j.heliyon.2023.e20242

Njock PGA, Yin ZY, Zhang N, 2025. High‐fidelity data augmentation for few‐shot learning in jet grout injection applications. *International Journal for Numerical and Analytical Methods in Geomechanics*, 49(1):83-100. https://doi.org/10.1002/nag.3862

Oliveira FM, Balbino MS, Zarate LE, et al., 2024. Predicting inmates misconduct using the shap approach. *Artificial Intelligence and Law*, 32(2):307–317. https://doi.org/10.1007/s10506-023-09352-z

Ribeiro MT, Singh S, Guestrin C, 2016. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, Association for Computing Machinery, p.1135-1144. https://doi.org/10.1145/2939672.2939778

Salehi S, Arashpour M, Golafshani EM, et al., 2023. Prediction of rheological properties and ageing performance of recycled plastic modified bitumen using machine learning models. *Construction and Building Materials*, 401 https://doi.org/10.1016/j.conbuildmat.2023.132728

Sang XK, Xu LJ, 2022. Research on the generation of creative animation driven by deep learning model. *Scientific Programming*, 2022(1) https://doi.org/10.1155/2022/5815693

Seyyedattar M, Zendehboudi S, Butt S, 2022. Relative permeability modeling using extra trees, anfis, and hybrid lssvm–csa methods. *Natural Resources Research*, 31(1):571-600. https://doi.org/10.1007/s11053-021-09950-1

Shapley LS, 1952. A value for n-person games. In: Kuhn, H., Tucker, A. Eds.), Contributions to the theory of games ii. Princeton University Press, Princeton, p.307–317.

Sun DL, Wu XQ, Wen HJ, et al., 2023a. A lightgbm-based landslide susceptibility model considering the uncertainty of non-landslide samples. *Geomatics Natural Hazards & Risk*, 14(1) https://doi.org/10.1080/19475705.2023.2213807

Sun YL, Dong YA, Wang DH, et al., 2023b. Correlation between travel experiences and post-covid outbound tourism intention: A case study from china. *Journal of Zhejiang University-SCIENCE A*, 24(11):1003-1016. https://doi.org/10.1631/jzus.A2300057

Tsaregorodtsev A, Belagiannis V, 2023. Particleaugment: Sampling-based data augmentation. *Computer Vision and Image Understanding*, 228 https://doi.org/10.1016/j.cviu.2023.103633

Wang DN, Li L, Zhao D, 2022. Corporate finance risk prediction based on lightgbm. *Information Sciences*, 602:259-268. https://doi.org/10.1016/j.ins.2022.04.058

Wang Y, Wang T, 2020. Application of improved lightgbm model in blood glucose prediction. *Applied Sciences*, 10(9) https://doi.org/10.3390/app10093227

Wen LY, Zhang XM, Li QF, et al., 2023. Kga: Integrating kpca and gan for microbial data augmentation. *International Journal of Machine Learning and Cybernetics*, 14(4):1427-1444. https://doi.org/10.1007/s13042-022-01707-3

Unedited

Wrzesinski G, Markiewicz A, 2022. Prediction of permeability coefficient k in sandy soils using ann. *Sustainability*, 14(11)
https://doi.org/10.3390/su14116736

Xing LM, Zhang YJ, 2022. Forecasting crude oil prices with shrinkage methods: Can nonconvex penalty and huber loss help? *Energy Economics*, 110
https://doi.org/10.1016/j.eneco.2022.106014

Yang Y, Zhou H, Wu JR, et al., 2022. Robustified extreme learning machine regression with applications in outlier-blended wind-speed forecasting. *Applied Soft Computing*, 122:108814.
https://doi.org/10.1016/j.asoc.2022.108814

You MY, Lu AN, 2021. A robust tdoa based solution for source location using mixed huber loss. *Journal of Systems Engineering and Electronics*, 32(6):1375-1380.
https://doi.org/10.23919/jsee.2021.000117

Zhang N, Xu KP, Yin ZY, et al., 2025. Finite element-integrated neural network framework for elastic and elastoplastic solids. *Computer Methods in Applied Mechanics and Engineering*, 433
https://doi.org/10.1016/j.cma.2024.117474

Zhao W, Yin QG, Wen LF, 2023. Intelligent inversion analysis of hydraulic engineering geological permeability coefficient based on an rf–hho model. *Water*, 15(11)
https://doi.org/10.3390/w15111993

Zhong DH, Liu DH, Cui B, 2011. Real-time compaction quality monitoring of high core rockfill dam. *Science China Technological Sciences*, 54(7):1906-1913.
https://doi.org/10.1007/s11431-011-4429-6

Zhou SX, 2023. An analysis of the small sample datasets based on machine learning. Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering, Association for Computing Machinery, p.1654–1658.
https://doi.org/10.1145/3573428.3573720

中文概要

题　目：小样本条件下基于改进 GAN 和可解释集成学习的土石料渗透系数预测

作　者：俞承禹，余红玲，瞿晓峰，刘宝熙，徐良嗣，刘新宇，陈翔宇

机　构：中国农业大学，水利与土木工程学院，中国北京，100083

目　的：本文旨在针对土石料渗透系数预测中存在小样本和可解释性不足的问题，探究基于生成模型的数据增强方法，以提高预测模型在小样本条件下的泛化能力。同时，结合具备可解释性的集成学习算法，增强预测结果的可信度，实现对土石料渗透系数的高精度预测。

创新点：1. 提出了一种基于改进生成对抗网络的数据增强方法，有效提升了小样本条件下渗透系数预测模型的性能；2. 构建了基于改进 LightGBM 的渗透系数预测模型，结合优化算法实现了渗透系数更高精度的预测；3.使用 SHAP 方法对预测结果进行全局和局部解释，增强了模型的可解释性；4. 将所提方法应用于实际土方工程案例中，验证了所提方法在工程实践中的有效性。

方　法：1. 将 Wasserstein 距离作为损失函数引入到条件生成对抗网络中，提出了 Wasserstein 条件生成对抗网络，以生成高质量的数据；2. 利用 LightGBM 算法建立了具有 Huber 损失函数和鱼鹰优化算法的高精度渗透系数预测模型；3. 使用 SHAP 方法，探究影响预测结果的关键特征，分析不同特征在数据集中的具体作用。

结　论：1. 基于 Wasserstein 条件生成对抗网络的数据增强方法能够生成高质量的样本，有效解决了小样本数据问题；2. 基于结合 Huber 损失和鱼鹰优化的 LightGBM 算法建立的渗透系数预测模型具有较高的预测性能；3.使用 SHAP 方法能够对预测结果进行全局和局部分析，提升预测模型的可解释性。

关键词：渗透系数预测；LightGBM；Wasserstein 条件生成对抗网络；SHAP

Unedited