Clarity in DEM cementation predictions: Integrating automated machine learning and interpretability analysis of shear wave velocity

Meng Sun[1]; Dong Liang[2]; Jiajia Wang[3]; Bate Bate[4]; Fan Xue[5]*

Abstract

Although shear wave velocity (Vs) is a fundamental parameter for evaluating the small-strain shear stiffness of materials, conventional discrete element method (DEM) simulations for predicting Vs in cementation analysis are computationally intensive. While machine learning (ML) provides a promising alternative for developing high-performance predictive models, the inherent opacity of most ML models diminishes their credibility, limiting practical applications. This paper introduces an approach that integrates automated machine learning (AutoML) with interpretability analysis. This approach achieves high-accuracy Vs prediction ($R^2 = 0.952$) and offers physically meaningful insights through integrated global-local interpretability. Analysis of 1,972 DEM samples with 10 microscopic features, the approach confirms the significant influence of features such as *ShortestDistance* and *HertzNumber* on Vs evolution. The contribution of this paper is twofold. First, the improved AutoML model marks a breakthrough by enabling more accurate Vs prediction in DEM cementation simulations while greatly reducing computational costs. Second, the interpretability results improve usability by unveiling transparent, physically grounded relationships between microscopic features and the macroscopic property Vs. The developed approach can enhance practical reinforcement applications with greater credibility, especially in quality control and process optimization.

[1] Meng Sun, PhD student
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong, China
E-mail: sunmhku01@connect.hku.hk ORCiD: https://orcid.org/0000-0002-7141-2833
[2] Dong Liang, PhD candidate
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong, China
E-mail: leodong@connect.hku.hk ORCiD: https://orcid.org/0000-0002-2918-7898
[3] Jiajia Wang, PhD candidate
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong, China
E-mail: wjiajia@connect.hku.hk ORCiD: https://orcid.org/0000-0003-3630-2003
[4] Bate Bate, PhD, Associate Professor
Institute of Geotechnical Engineering, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, 310058, China
E-mail: batebate@zju.edu.cn ORCiD: https://orcid.org/0000-0002-8692-8402
[5] Fan Xue, PhD, Associate Professor
Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong, China
National Center of Technology Innovation for Digital Construction Hong Kong Branch, The University of Hong Kong, Pokfulam, Hong Kong, China
E-mail: xuef@hku.hk ORCiD: https://orcid.org/0000-0003-2217-3693
* Corresponding author, Tel: +852 3917 4174, Fax: +852 2559 9457, Email: xuef@hku.hk

Keywords

Automated machine learning (AutoML); interpretability analysis; Explainable AI (XAI); discrete element method (DEM); cementation simulation; shear wave velocity (Vs)

Highlights

· Automated ML of discrete element method (DEM) achieves robust Vs prediction (R^2 = 0.952)
· Interpretability pipeline (PDP, ALE, LIME, SHAP) resolves ML black-box concerns
· Key microscopic features *ShortestDistance*, *HertzNumber* drive Vs evolution trends
· Our approach outperforms $C_{d,n}$-based fitting with higher accuracy and generalization
· Transparent model enables broader microscopic studies and practical applications

Nomenclature

Abbreviation	Explanation
AEC	Architecture, Engineering, and Construction
AI	Artificial intelligence
ALE	Accumulated Local Effects
AutoML	Automated machine learning
$C_{d,n}$	The normalized connectivity per unit distance
DEM	Discrete element method
G_0	Shear stiffness
ICE	Individual Conditional Expectation
K_0	Static lateral stress
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large language model
LPB	Linear parallel bond
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MC	Monte-Carlo
ML	Machine learning
MSE	Mean Squared Error
NDT	Non-destructive testing
PDP	Partial Dependence Plot
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
RNN	Recurrent neural networks
SHAP	SHapley Additive exPlanations
Vs	Shear wave velocity
XAI	Explainable AI
ρ	Sample density

2

1 Introduction

Recent advances in microscopic mechanical studies have promoted the widespread use of material cementation reinforcement in the Architecture, Engineering, and Construction (AEC) industry (Ashraf et al. 2017). Shear wave velocity (Vs) is essential because it directly relates to the small-strain shear stiffness (G_0), a key parameter for assessing the effect of cementation-based reinforcement. While numerical simulations like the discrete element method (DEM) can effectively solve static problems, dynamic simulations—especially those involving Vs prediction during cementation—are computationally demanding and time-consuming (Sun et al. 2025; Yang et al. 2017). As a result, understanding the microstructural influences on such dynamic small-strain behaviors remains difficult (Bate et al. 2021). Therefore, the AEC industry needs efficient solutions for predicting Vs and analyzing how microscopic features affect dynamic small-strain behavior.

These requirements for precise small-strain analysis indicate the necessity of incorporating machine learning (ML) into the AEC industry (Singh et al. 2021). ML can learn nonlinear relationships from high-dimensional data, enabling valuable analysis and identification of microscopic mechanical behavior (Zhang et al. 2024). Modern ML techniques, especially supervised learning algorithms, have proven effective in solving inverse problems and predicting mechanical properties without depending on physics-based models (Deka 2019; Liang & Xue 2023). Successful solutions in this area often combine domain expertise, experimental methods, and numerical simulations with advanced ML technologies (Saka et al. 2023; Sun et al. 2024). Furthermore, common optimization techniques such as data-level approximate computing, precision scaling, and feature selection enhance ML models (Dalloo & Humaidi 2024; Euldji et al. 2023). Recent research has adopted ML algorithms and hybrid techniques for Vs prediction, indicating innovative integrations of physical research in the AEC industry (Ghorbani et al. 2025; Hazbeh et al. 2024; Makarian et al. 2023; Rajabi et al. 2023).

Automated ML (AutoML) has recently gained attention for solving complex prediction tasks in simulations. AutoML aims to automate all stages of ML, including data preprocessing, feature engineering, model selection, hyperparameter tuning, and deployment. Thus, non-experts are enabled to develop high-performance models directly from provided datasets to solve the target tasks (Baratchi et al. 2024; Barbudo et al. 2023). However, AutoML still faces challenges due to the "black-box" nature of many ML models, hindering its widespread application.

Interpretability analysis improves the transparency and credibility of complex models, which is significant in the era of "Artificial Intelligence (AI) for Science" (Zhang et al. 2023). The implementation of GDPR has introduced requirements for trustworthy data processing, making interpretability a key research area in AI and ML (Varshney 2016). Explainable Artificial Intelligence (XAI) is truly driving the development of AI technology towards greater transparency, credibility, and responsibility. Currently, interpretability methods examine

complex models through visualization, model simplification, feature-based interpretation, and rule-based interpretation (Bach et al. 2015; Barredo Arrieta et al. 2020).

This paper presents the application of AutoML and interpretability analysis for predicting Vs in DEM cementation simulations to address the computational cost of conventional DEM and the lack of credibility in ML models. A dataset of 1,972 DEM samples with measured Vs values was analyzed. AutoML was employed to predict Vs based on physically induced microscopic features. Two global interpretability methods, i.e., PDP and ALE, and two local interpretability methods, i.e., LIME and SHAP, were used to provide reasonable microscopic interpretations at both global and local scales. The findings demonstrate substantial advances in prediction performance and offer effective interpretations of microscopic features. Section 2 reviews related research. Section 3 details the proposed research methods. Section 4 presents and analyzes the experimental results. Section 5 discusses the implications of the findings, and Section 6 concludes the paper. The contributions of this paper are twofold. First, the improved Vs prediction leads to higher accuracy in DEM cementation simulations while simultaneously reducing computational cost. Second, the interpretability analysis enhances prediction credibility and can provide valuable guidance for practical reinforcement applications.

2 Related research

2.1 Cementation content and Vs in soil reinforcement

Cementation simulation is a widely adopted method (Ashraf et al. 2017; Zhao et al. 2022). This method effectively examines materials' reinforcement processes within the AEC industry (DeJong et al. 2010; Kempfert & Gebreselassie 2006). Cementation reinforcement affects the stress-strain relationship of materials and improves their shear stiffness (Karol 2003). Various cementation methods are widely applied to enhance soil properties and address instability in soft soils (Bang et al. 2009; Martinez & DeJong 2009; Salifu et al. 2016). However, the uncontrollable nature of cementation processes often leads to non-uniform reinforcement, which creates engineering risks (Bate et al. 2021; Niu & Zhang 2018). This necessitates further investigation into the microscopic mechanisms involved.

Shear wave velocity (Vs) is used to assess the effect of reinforcement because it is a reliable non-destructive testing (NDT) method used to measure shear stiffness—the material's resistance to deformation under shear stress (Martinez et al. 2013; Hu & Wang 2024; Han et al. 1986). The shear stiffness (G_0) is calculated by the equation:

$$G_0 = \rho \mathrm{V}_s^2, \tag{1}$$

where ρ is the sample density, and Vs is the shear wave velocity. Vs, as a dependable metric, assesses the reinforcement effects in cementation simulations, particularly for strengthening loose sandy soils. Vs exhibits a nonlinear positive correlation with added cement (cementation content). DeJong et al. (2010) compared the performance of various NDT methods, such as shear wave, compression wave, and resistivity, and showed the effectiveness of Vs in evaluation. The relationship between Vs and cementation content was also preliminarily

characterized (Martinez et al. 2013). Sun et al. (2022) performed a discrete element method (DEM) simulation using macro parameters and created a microscopic digital twin model of cemented samples to quantify the evolution of Vs; experiments in other studies verified this quantification process. Based on geotechnical engineering principles, these investigations reveal Vs evolution at elevated cementation levels and offer a microscopic interpretation of Vs.

Various methods have been proposed to predict or estimate Vs, including empirical corrections, numerical simulations, and ML approaches (Hazbeh et al. 2024). Table 1 provides typical categories of these Vs prediction methods.

Table 1 Summary of Vs prediction methods

Method category	Advantages	Limitations	Representative methods
Empirical correlation	Low computational cost, physical insights, and high interpretability	Failing to represent complex nonlinear relationships, poor accuracy	Vs-DCT relationship (Han et al. 1986), Vs-m_n fitting (Bate et al. 2021), GM estimation (Makarian et al. 2023)
Optimization-based algorithms	Strong theoretical foundation and high interpretability, handling uncertainty	High computation cost, convergence issues, indirect prediction	Cuckoo Optimization Algorithm (Anemangely et al. 2019), Monte-Carlo simulation (Sun et al. 2022)
Classical machine learning methods	End-to-end prediction, capturing complex nonlinear relationships, relatively low cost	Strong dependence on data and feature quality, limited physical consistency	SVM (Anemangely et al. 2019), Random forest (Hazbeh et al. 2024)
Deep learning methods	End-to-end prediction, high accuracy	Black-box nature, massive data, high computational cost	DBN (Hazbeh et al. 2024), Hybrid CNN&LSTM (Ghorbani et al. 2025)
DEM-simulation-based methods	Physical insights and high interpretability, real-world simulation	high computational cost, parameter calibration challenge	Granular material simulation (Sadd et al. 2000), CaCO3 cementation prediction (Sun et al. 2025)

The measurement is costly and time-consuming, and ML methods offer alternative solutions but often lack interpretability. DEM simulation has become a key method for predicting Vs in cementation analysis. However, constructing a DEM model to simulate shear wave propagation is complex and challenging (Table 1) (Sun et al. 2022). This growing complexity arises from the increasing parameters, difficulties in timestep selection, the complex response to nonlinear material behavior, and complex dynamic processes. These challenges make small-strain analysis, like Vs prediction and microscopic feature effects, considerably more complex than constructing the initial sample.

2.2 Preliminary applications of AutoML and interpretability

AutoML enables users to benchmark and integrate various algorithms in rapidly developing high-performance ensemble models (Li et al. 2023). Compared with traditional ML, AutoML simplifies the workflow for feature engineering, model selection, and parameter tuning. Specifically, automated Bayesian optimization and model weight assignment relieve researchers of the burden imposed by redundant and complex parameter tuning (Barbudo et al.

5

2023). Moreover, AutoML allows non-experts to construct high-performance pipelines satisfying specific applications (De Bie et al. 2022). Auto-sklearn, an increasingly prominent AutoML toolkit, integrates a diverse set of algorithms, feature preprocessing methods, and data preprocessing methods. In contrast to traditional machine learning algorithms, AutoML reconfigures these processes as hyperparameters (Feurer et al. 2019; 2022). The efficiency of AutoML has captured the interest of researchers in the AEC industry. The failure mechanism of reinforced concrete shear walls can be effectively classified by input design parameters (Liang & Xue 2023), and high-performance non-destructive concrete strength prediction is also possible (Sun et al. 2023).

Although machine learning algorithms perform well in numerous tasks, many engineers and practitioners remain hesitant to use them due to their black-box nature (Lundberg & Lee 2017; Zhao & Hastie 2021). Model trustworthiness and decision causality require interpretability for effective risk control. Interpretability refers to the degree to which the rationale behind a decision is understandable (Kim et al. 2016; Miller 2019; Molnar 2020). The black-box nature of algorithms used to train AutoML models verifies the importance of Explainable AI (XAI), particularly model-agnostic interpretability methods. Therefore, these methods warrant significant attention from researchers.

Partial Dependence Plot (PDP) and Accumulated Local Effects (ALE) are two typical global interpretability methods (Apley & Zhu 2020; Friedman 2001), whereas Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are two typical local methods (Lundberg & Lee 2017; Marcílio & Eler 2020; Ribeiro et al. 2016a; 2018). The summary of typical interpretability methods is presented in Table 2. In the AEC industry, researchers employed SHAP to interpret both global and local concrete strength predictions, using feature contribution visualizations to inform concrete mixing decisions (Sun et al. 2023). SHAP was also applied to quantify the contribution of particle size distribution to shear stiffness in DEM simulations (Liu et al. 2024). Zhang et al. (2025) used SHAP and PDP to analyze the influence of four factors in asphalt preparation on predicting dynamic modulus. However, alternative interpretability methods have limited applications within the AEC domain. This scarcity presents an opportunity for further exploration.

Table 2 Summary of typical model-agnostic interpretability methods (Apley & Zhu 2020; Friedman 2001; Lundberg & Lee 2017; Marcílio & Eler 2020; Ribeiro et al. 2018)

Method	Context	Limitations
PDP	Global interpretation, visualizing the average marginal effect of a feature	Feature independence assumption
ALE	Global interpretation, correlated features handling, unbiased estimation.	Less intuitive than PDP, complex interpretation plot
LIME	Local interpretation, local fidelity	Dependent on perturbation parameters, sensitive to kernel width
SHAP	both local and global interpretation	High-cost computation

2.3 Cementation modeling via DEM

DEM has become a key method in cementation simulations because of its capacity to model each material particle at the microscopic scale (Lin et al. 2016; Sun et al. 2022). This capability is vital, as it allows for detailed analysis of soil, structures, and reinforcement effects in infrastructure engineering. DEM can control the state of each particle via coating and pore filling to simulate the global cementation content (Niu et al. 2024; Xu et al. 2024). With advances in computational power, these simulation capabilities make DEM an effective alternative to in-situ practices and experimental observations (DeJong et al. 2010; Jiang 2019; Martinez et al. 2013). Chang et al. (2016) summarized biomineralization treatment methods for loose sand and proposed specific microscopic models. Yun and Evans (2011) used DEM to measure the static lateral stress (K_0) and Vs of cemented sand. Yang et al. (2017) conducted a 3D DEM analysis of reinforced sand under drainage triaxial compression. Feng et al. (2017) simulated macroscopic and microscopic responses before and after cementation.

Sun et al. (2025) simulated numerous DEM samples with measured Vs to investigate the spatial variability of cementation distribution. In this paper, the DEM simulation of basic sand sample cementation proceeded as follows: particles represented the sand; Hertz contact modeled interactions between uncemented particles, and linear parallel bond (LPB) contact modeled interactions between cemented particles. The simulation began by applying the collected experimental macro parameters to calibrate the simulation parameters. Then, an uncemented DEM sample consisting only of Hertz contacts was generated to represent loose soft soil. Once this initial uncemented DEM sample was created, different DEM samples with specified cementation content were simulated by converting Hertz contacts to LPB contacts, based on the cementation content algorithm. This methodology enabled the simulation of shear wave propagation across diverse cementation contents to acquire Vs. Significantly, variations in cementation attributes—such as position, size, and clustering—produce differences in Vs. This occurs even when cementation content remains identical.

Fig. 1 illustrates a typical sample and propagation process under 2.55% cementation content. In the DEM model, particles had the same properties, while the distribution of LPB contacts influenced wave propagation and Vs. Previous microscopic studies of these DEM samples indicated that the preferential path of shear wave propagation determines Vs. Nevertheless, two critical questions remain unanswered. First, a method for directly determining Vs for individual samples remains unclear. Second, the specific microscopic features influencing Vs along the propagation path need interpretation.
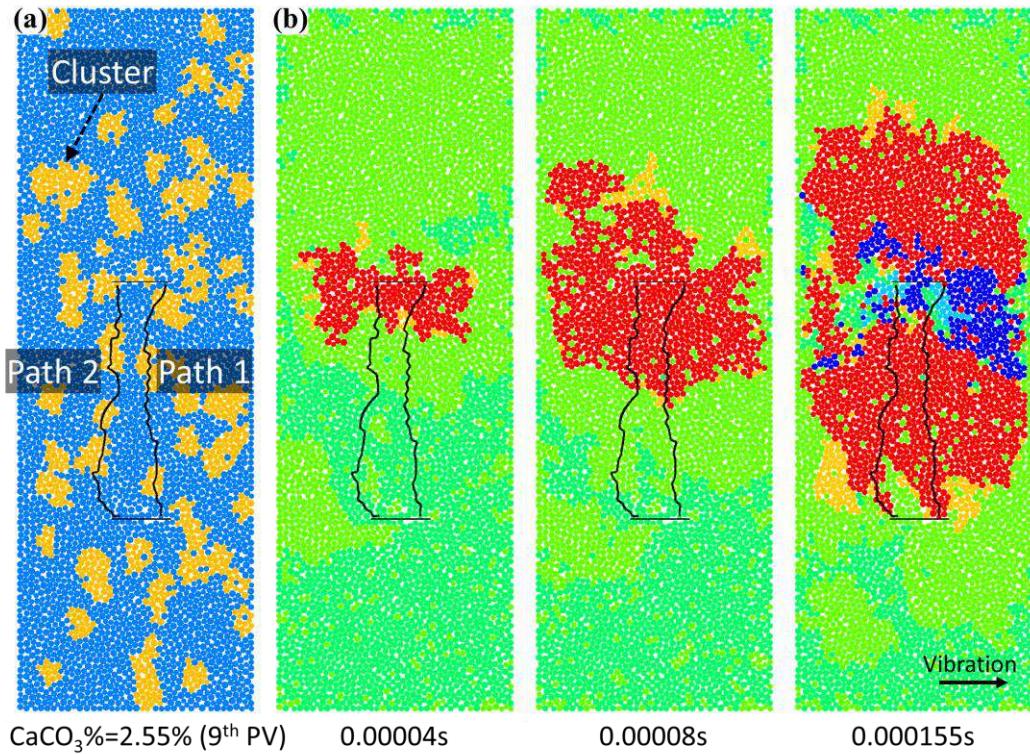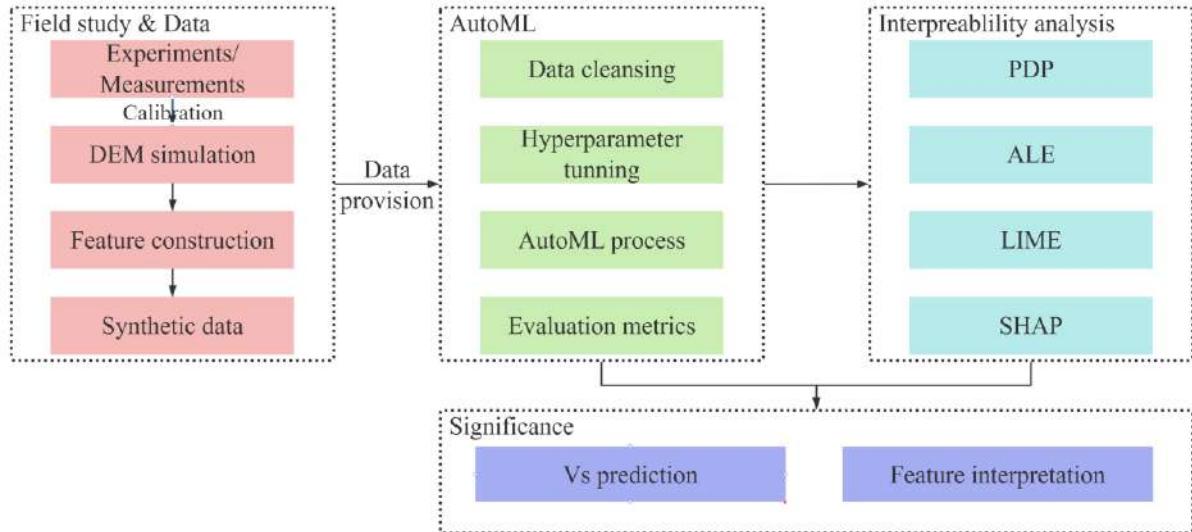
Fig. 1 (a) A typical sample under the cementation content of 2.55% (Blue denotes pure sands, orange denotes cemented clusters, and black lines denote the shear wave propagation path); (b) Wave propagation process (green indicates sands without vibration, red and blue indicate sands with vibration) (Sun et al. 2022).

To address these challenges, researchers need efficient methodologies to predict Vs and clarify microscopic features. Conventional DEM simulation relies on manual calibration and iterative processes, which are time-consuming and inefficient. Given that prior Vs measurements in cementation samples provide many microscopic features, AutoML-based end-to-end prediction represents a feasible and promising research direction (Sun et al. 2025). Interpretability analysis enables the investigation of dominant microscopic features. Therefore, the integration of AutoML and interpretability techniques in Vs prediction is promising, as this combination enhances both predictive accuracy and mechanistic insight.

3 Research methods

To conduct the Vs prediction and microscopic interpretation from the DEM model, we propose an approach that integrates AutoML with interpretability analysis. In this Section, Fig. 2 presents the workflow, including the case dataset derived from DEM simulations, the AutoML regression pipeline for Vs prediction, and four interpretability methods. The pseudocode of this approach is presented in Algorithm S1 in Appendix D.

Fig. 2 The conceptual diagram for Vs prediction and microscopically interpretability analysis.

3.1 Case dataset

This paper uses a total of 1,972 DEM samples with measured Vs published by Sun et al. (2025) to construct a dataset of trusted synthetic data. The procedure for acquiring the sample data is provided in Appendix B of the supplementary material. The dataset spans a wide range of cementation content levels. On a microscopic scale, the propagation of shear waves results from the transfer of force between particles through contact. Vs is dominated by the preferential propagation path of shear waves (Bate et al. 2021; Santamarina et al. 2001; Sun et al. 2022). Ten typical physical features of the contact along the propagation path were extracted to quantify their detailed contributions to Vs. The definition and distribution of these features are illustrated in Table 3 and Fig. S1 (Appendix C). In the subsequent sections, this paper focuses on the relationship between the extracted microscopic features and Vs and leaves the macroscopic quantity of cementation content aside.

Table 3 The definition of 10 microscopic features

ID	Features / Label	Definition
F_1	LPBPercent	The proportion of the number of LPB contacts to all the contacts on the shortest path
F_2	ContactsNumber	The quantity of all contacts on the shortest path
F_3	LPBConstantMaxNumber	The quantity of LPB contacts within the longest consecutive LPB contacts on the shortest path
F_4	split0	The quantity of segments with LPB contacts split by Hertz contacts on the propagation preference path
F_5	split1	The quantity of segments with more than one LPB contact split by Hertz contacts on the propagation preference path
F_6	LPBNumber	The quantity of LPB contacts on the shortest path
F_7	HertzNumber	The quantity of hertz contacts on the shortest path
F_8	LPBconstantPrct	The proportion of the number of LPB contacts within the longest consecutive LPB contacts to the total number of contacts on the shortest path
F_9	ShortestDistance	Weighted shortest distance from the transmitter to the receiver
F_{10}	averageSplitLen	The average length of split segments with LPB contacts
	Vs	Shear wave velocity

9

3.2 AutoML regression

This paper applies Auto-sklearn to predict Vs based on DEM simulations. Fig. 3 shows the AutoML regression process. Microscopic features are extracted from the DEM simulation to construct the trusted synthetic dataset. Then, a 10-fold cross-validation procedure is employed to select models and mitigate bias. This is an iterative process in which nine folds are used for training, and one fold is designated for validation (hold-out set). To prevent data leakage, all DEM samples generated from the same initial spatial configuration are grouped together and assigned to the same fold.

Data cleansing	Encoding methods	Feature preprocessors	Regressors		Evaluation metrics
N/A completion	Standard scale	No preprocessing	Bayesian Ridge Regression	Orthogonal Matching Pursuit	MSE
	Pass through	Extra trees preproc for regression	Decision Tree	Passive Aggressive	MAE
		FastICA	Elastic Net	Random Forest	RMSE
		Feature agglomeration	Extra Trees	SGD	R^2
		Kernel PCA	Gradient Boosting	Adaboost	
		Kitchen sinks	Kernel Ridge Regression	Ard regression	
		PCA	K Nearest Neighbors	Gasusian process	
		Polynomial	Lasso	Support Vector Regression	
		Random trees embedding	MLP		

Fig. 3 Structural schematic and process details in the Auto-sklearn prediction.

Within the Auto-sklearn system, hyperparameter tuning is automated to maximize the automation advantages of AutoML, except that configuration parameters must be preset. Prior to applying each individual model, meta-learning is employed to optimize hyperparameters by calculating the similarity between the training dataset and existing datasets (Feurer et al. 2015).

During each model training, all samples in the dataset were normalized and encoded using either standard scaling or pass-through methods for the 10 input features (Fig. 3). Namely, two continuous features, *ShortestDistance* and *averageSplitLen*, were rescaled to the range [−1, 1] by subtracting the mean and scaling to unit variance. The remaining features were left untransformed. This encoding step is essential to prevent the adverse effects of varying data scale. After the encoding step, a collection of 11 feature preprocessing methods was used,

including polynomial features and feature agglomeration. Then, 17 regressors, including *K* Nearest Neighbors (Huang et al. 2019), Gaussian Processes (Rasmussen et al. 2004), and Extra Trees (Geurts et al. 2006), are selected to predict Vs both independently and collectively. Collective predictions are made using ensemble methods, including *Ensemble Selection*, *Library Pruning*, and *SingleBest*. The final prediction is evaluated using metrics such as MSE (Mean Squared Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 (Coefficient of determination) to ensure the reliability and performance of the ensemble model. Examples of prediction outputs are listed in Table 4.

Table 4 Examples of features and predicted values

Instance Id	Features for prediction					Vs (m/s)	
	F_1	F_2	F_3	…	F_{10}	predicted	DEM measured
1	0	41	0	…	-0.753	195.42	195.44
2	0.073	41	1	…	-0.644	206.08	209.52
3	0.262	42	1	…	-0.644	241.87	232.09
…	…	…	…	…	…	…	…
1,972	0.950	40	38	…	-1.597	2105.74	2090.84

3.3 Interpretability methods

3.3.1 Partial dependent plot (PDP)

A Partial Dependence Plot (PDP) is a global method used to describe the marginal effect of each feature on the predicted outcome. PDP assumes that the first feature under study is unrelated to the second feature (if applicable). The result arises from observing changes in the predicted outcome after shuffling the values of the feature in question. This method quantifies the contribution of the feature to the prediction. PDP provides a causal interpretation that is intuitive and implementable (Zhao & Hastie 2021).

The PDP analysis process is as follows (using the feature F_1 as an example):
   a. Shuffle the F_1 values within the dataset while fixing all other feature values.
   b. Calculate the average output for different values of F_1 under Monte Carlo conditions.
   c. Plot the PDP result, showing the average curve.

To supplement PDP, Individual Conditional Expectation (ICE) provides an individual sample perspective and enhances the understanding of the impact of features. The combined use of PDP, which gives an overall view of feature influence, and ICE, which provides individual sample insights, enables a deeper understanding of the model's behavior and the role of features.

3.3.2 Accumulated local effects (ALE)

The Accumulated Local Effects (ALE) method is a distinct interpretable method that effectively addresses feature correlation issues. ALE calculates the variance, accumulation, and centralization of predicted values. Features are partitioned into multiple bins to ensure the points within each bin reflect the actual data points. For each instance within a bin, the

11

differences in predicted values result from substituting the feature with both the upper and lower bounds of the bin. These differences are subsequently accumulated and centered to form the ALE curve.

3.3.3 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) is a locally additive feature attribution method (Ribeiro et al. 2016b) commonly used in classification tasks. In the LIME process, a new dataset (local samples) is generated by perturbing the original data; the original predicted value serves as the reference. An interpretable model (e.g., linear model, decision tree) trains using the local samples. This interpretable model then provides a data representation with features and allows observation of the importance of each feature. This paper treats the regression task as a matrix to execute data perturbation through sampling and inverse operations.

The detailed process is as follows (using feature F_1 as an example).

  a. Train the uninterpretable AutoML model and select the feature F_1 to interpret.
  b. Generate N perturbations of the data to generate local samples.
  c. Calculate the weight of the samples, which corresponds to the distance between the perturbed data and the feature to be interpreted.
  d. Fit a new interpretable model based on the new dataset.
  e. Interpret the contribution of feature F_1 in the AutoML model using the interpretable model.

3.3.4 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) is a method designed to interpret predictions of complex models in a unified framework (Lundberg & Lee 2017). As a local model-agnostic interpretable method, SHAP derives from game theory and quantifies the contribution of features to predictions for individual instances. The core of the SHAP method is the Shapley value, which provides an attribution value for each input feature. The Shapley value is additive and locally accurate. In this context, the predicted value can be interpreted as the sum of the Shapley values for all features and the global predicted average. The significance of SHAP lies in its ability to measure feature contributions consistently, which ensures a fair and consistent interpretation of each feature's impact on the prediction.

In summary, PDP intuitively interprets microscopic features and reveals interactions between features. ALE addresses the issue of feature correlation in global interpretations. LIME and SHAP focus on the interpretation of individual sample differences. LIME ensures local fidelity through 'faithful approximations', and SHAP, available globally and locally, offers further insight into the evolution of Vs for each sample.

4 Experiment results

4.1 Experimental Settings

The case dataset was used as the input for AutoML in Auto-sklearn (ver. 0.15). 10-fold cross-validation was employed to select models and reduce bias, enhancing model generalization (Fushiki 2011; Liang & Xue 2023). The time limits for automated model search, single model fitting, and memory limit were set to 300 seconds, 30 seconds, and 3,072 MB, respectively.

After training, the ensemble selection method automatically selected the best model from all the trained models and assigned weights based on performance. Additionally, the R^2 metric was used to evaluate the performance of the ensemble model. Four open-source tools—PDPbox (ver. 0.3), Scikit-Explain (ver. 0.1), Lime (ver. 0.2), and SHAP (ver. 0.44) —were selected to perform interpretability analysis for PDP, ALE, LIME, and SHAP, respectively. Based on fitting Vs versus the normalized connectivity per unit distance ($C_{d,n}$), two methods were selected as baseline (Sun et al. 2025).

4.2 Results of AutoML's Vs prediction
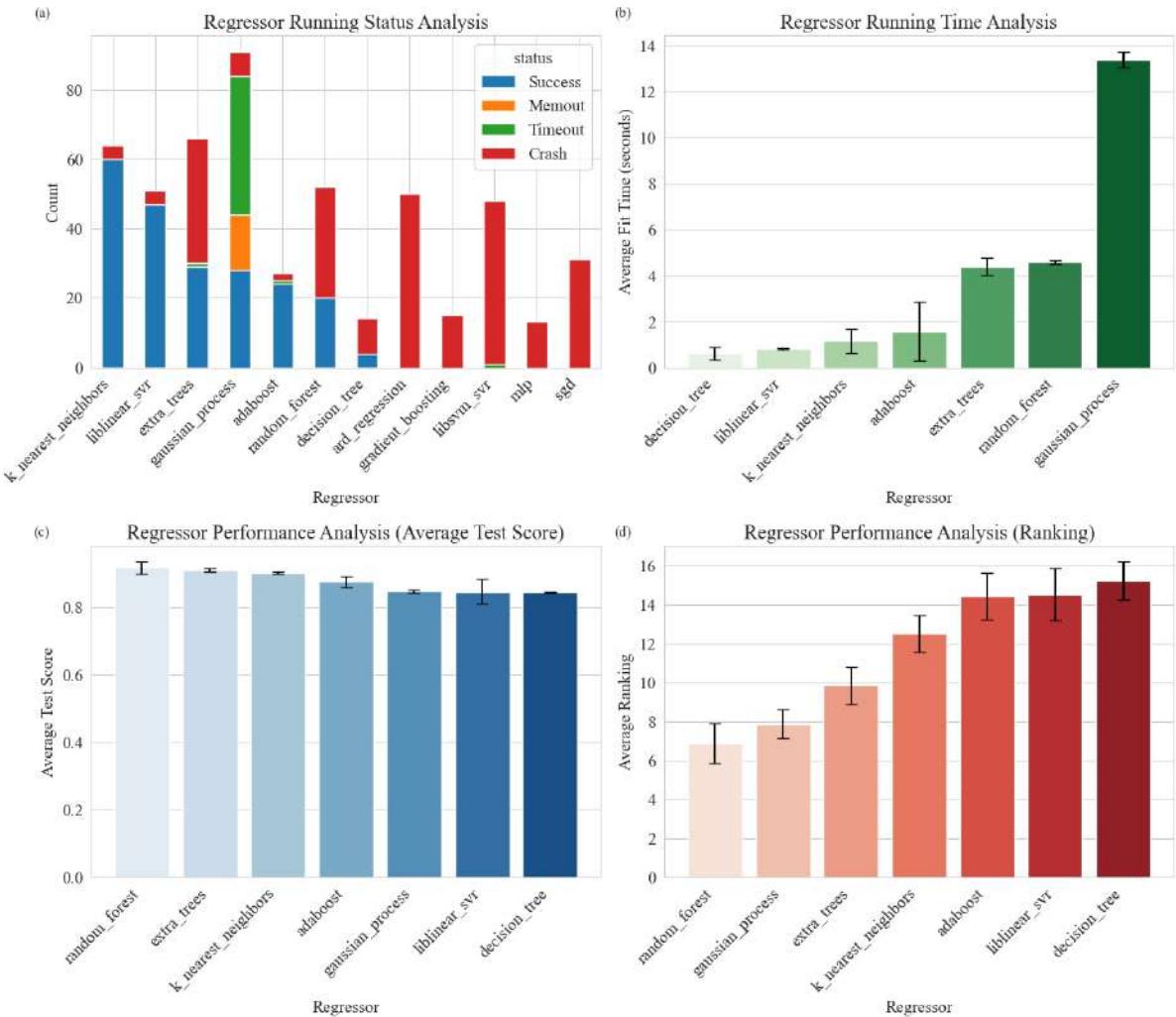
The AutoML process trained 1,972 samples and employed 17 types of machine learning regressors (Fig. 3). These regressors were used in a total of 66 models applied during the 10-fold training. The ensemble model was selected from these models, and its overall performance was outstanding, as shown in Table 5.

Table 5 Ensemble model performance in each fold

Fold	Best Ensemble Model	Accuracy (R^2)		
		Total	Train	Test
1	$0.34G_{1-1}+0.28G_{1-2}+0.34K_{1-1}+0.04A_{1-1}$	0.952	0.952	0.948
2	$0.5G_{2-1}+0.16G_{2-2}+0.1R_{2-1}+0.24K_{2-1}$	0.946	0.949	0.926
3	$0.32K_{3-1}+0.14G_{3-1}+0.12R_{3-1}+0.2L_{3-1}+0.16K_{3-2}+0.06K_{3-3}$	0.942	0.942	0.936
4	$0.82G_{4-1}+0.02R_{4-1}+0.16K_{4-1}$	0.936	0.940	0.902
5	$0.5G_{5-1}+0.14E_{5-1}+0.14R_{5-1}+0.12L_{5-1}+0.08K_{5-1}+0.02G_{5-2}$	0.944	0.944	0.943
6	$0.38G_{6-1}+0.1K_{6-1}+0.02E_{6-1}+0.18D_{6-1}+0.1L_{6-1}+0.12K_{6-2}+0.08E_{6-2}+0.02D_{6-2}$	0.943	0.945	0.923
7	$0.52G_{7-1}+0.1E_{7-1}+0.14K_{7-1}+0.02G_{7-2}+0.2K_{7-2}+0.02E_{7-2}$	0.941	0.942	0.937
8	$0.72G_{8-1}+0.08L_{8-1}+0.12G_{8-2}+0.06E_{8-1}+0.02G_{8-3}$	0.935	0.937	0.911
9	$0.46G_{9-1}+0.28G_{9-2}+0.1L_{9-1}+0.16K_{9-1}$	0.937	0.939	0.923
10	$0.24G_{10-1}+0.32G_{10-2}+0.04K_{10-1}+0.1L_{10-1}+0.02K_{10-2}+0.08K_{10-3}+0.1K_{10-4}+0.02E_{10-1}+0.02A_{10-1}+0.06L_{10-2}$	0.941	0.949	0.881

G: gaussian process; K: k nearest neighbors; A: adaboost; R:random forest; L: liblinear svr; E: extra_trees; D: decision_tree

Individual models should be investigated before applying the ensemble model. Fig. 4 summarizes the regressors using different algorithms in 522 single-model training sessions of the 10-fold cross-validation.

Fig. 4 Regressor analysis during the AutoML (10-fold) process.

The *K*-Nearest Neighbors and Liblinear_SVR were the top two regressors for the dataset, with more than 90% successful runs. In contrast, some regressors failed; this indicates their unsuitability for this Vs prediction task. Among the eight regressors that ran successfully, Fig. 4bcd shows the best performer for each regressor in each fold. This figure also presents test scores, rankings, runtime, and inter-fold errors analyses.

The Gaussian process achieved the highest score and ranked first in terms of test performance. However, it had a significantly longer execution time than the other regressors, with a total time nearly three times that of the others. Extra Trees also performed well in predictions, but its performance varied greatly across folds; this indicates poor robustness.

These regressors had distinct advantages and disadvantages, suggesting that an ensemble model would improve performance. Table 5 shows that the ensemble model for each fold was determined through post-hoc ensembling (Feurer et al. 2022). The evaluation metric R^2 was greater than 0.88 for all 10 folds in the test set. The ensemble size ranged from 3 to 10, and the Gaussian process formed the primary component in all 10 folds.

14

Fig. 5 presents the prediction performance metrics of the AutoML model compared to prior baseline methods. Both $C_{d,n}$-Coord and $C_{d,n}$-Monte-Carlo (MC) depended on connectivity among the preferential paths, discussed in prior studies (Sun et al. 2022; Sun et al. 2025). The AutoML prediction (RMSE = 143.472m/s) showed smaller errors compared to $C_{d,n}$-Coord (RMSE = 208.218 m/s) and $C_{d,n}$-MC methods (RMSE = 343.835 m/s). Within the 20% error interval, the AutoML approach achieved a higher percentage of predicted results (90.16%) than $C_{d,n}$-Coord (78.50%) and $C_{d,n}$-MC (44.27%) methods. Extreme outliers in the predictions of the $C_{d,n}$-related methods indicate incorrect predictions in some cases (dashed boxes in Fig. 5). In contrast, the AutoML predictions showed no excessive deviation; this demonstrated its robustness.
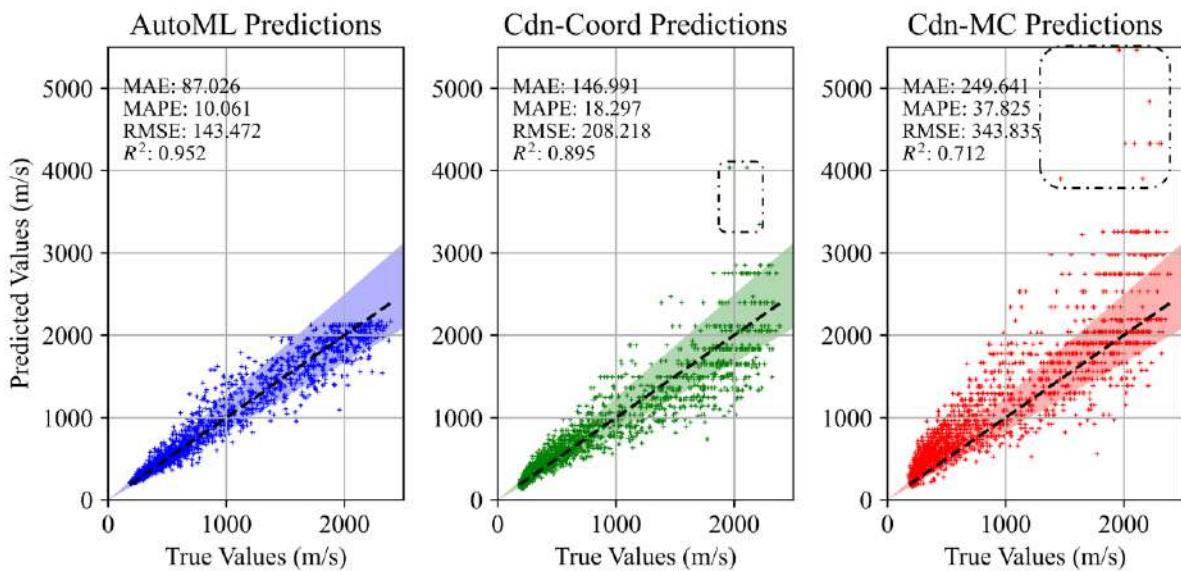
Fig. 5 Scatter plot of prediction results by the presented AutoML, $C_{d,n}$-Coord, and $C_{d,n}$-MC approaches, where the light color areas are 20% error intervals and dashed boxes highlight extreme outliers in predictions.

The AutoML approach reduced computational cost substantially compared to DEM-based simulations (Fig. 6). For a single Vs prediction task, repeated trial calibrations were indispensable in the DEM simulation, which was quite time-consuming. Plus, the Simulation Process 1 (843s) must be included in the prediction preparation. In contrast, the training process of AutoML only required 3048s. For each prediction in the task, the AutoML approach only costs 0.193s while the DEM-based simulation costs 337s.

Fig. 6 The computation time contrast between the DEM-based simulation and the AutoML approach in one Vs prediction task

425

4.3 Results of interpretability analysis

4.3.1 PDP & ICE

As a global interpretable method, Fig. 7 shows the thick blue PDP curves of the trained AutoML model, regarding the relationship between Vs (y-axis) and all 10 features (x-axis). Additionally,

430 light blue lines represent all individual conditional expectations (ICE).
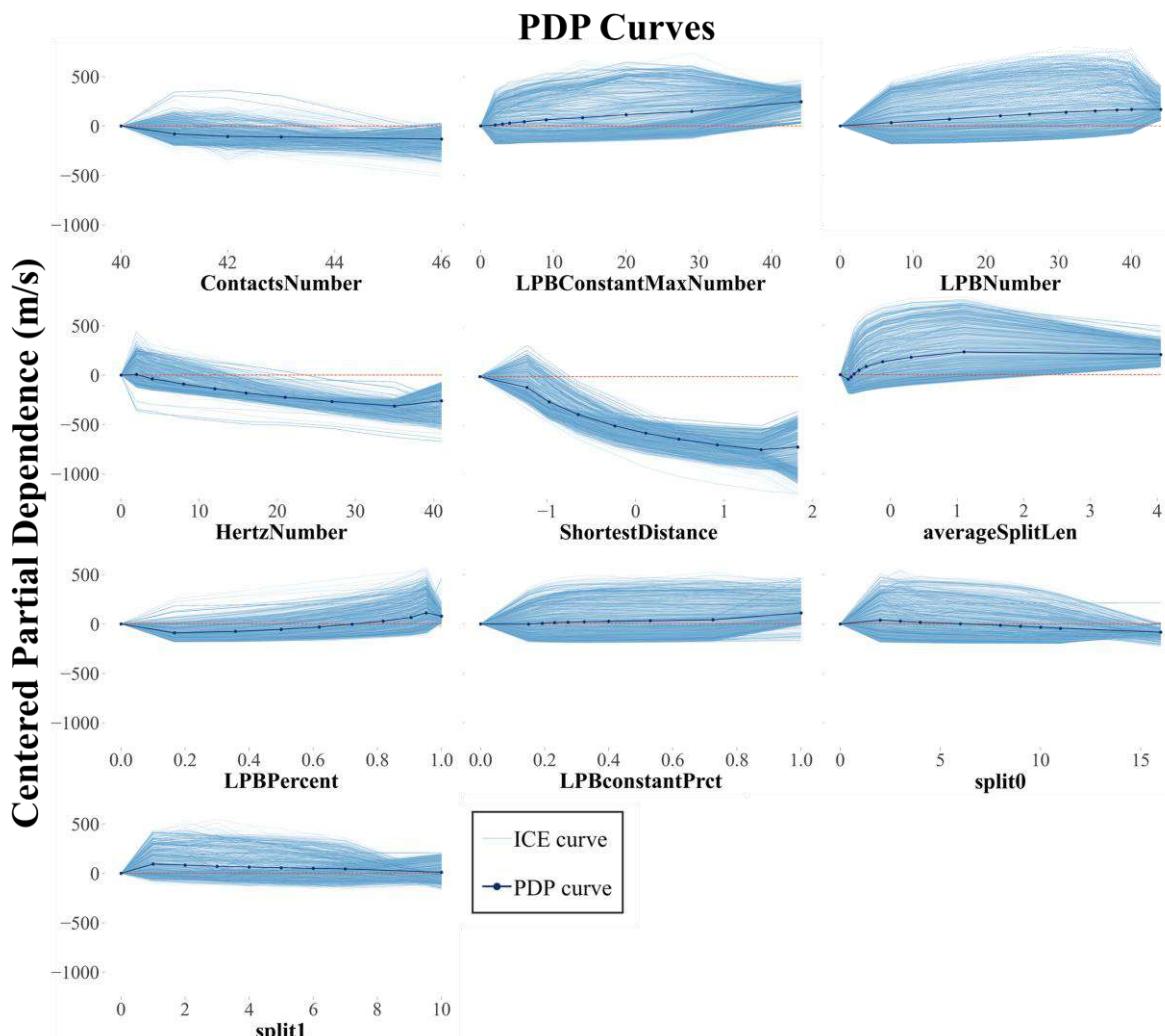
Fig. 7 PDP curves of 10 features regarding the trained AutoML model.

16

ShortestDistance is the most influential among all 10 features, as its maximum effect on Vs can be 750m/s. Given that this feature has attributes of contact types, propagation distance, and coordination number, the significance of *ShortestDistance* to Vs is understandable. The interpretability analysis result echoed prior studies that the preferential propagation dominates Vs (Bate et al., 2021; M. Sun et al., 2022). Additionally, the monotonically positive contribution of *LPBNumber* (170 m/s) and the negative contribution of *HertzNumber* (300 m/s) corroborated the microscopic analysis (Sun et al. 2022). This analysis stated that LPB contacts representing cemented particles provide strong interaction, and more LPB contacts/fewer Hertz contacts result in higher Vs.

However, due to the correlation between features, the accuracy of the interpretation was potentially limited, and the 2D PDP was used to indicate how the interaction of 2 related features influenced the prediction (Molnar, 2020). In this way, the inaccuracy caused by the feature independence assumption was addressed. All the interactions between feature pairs were visualized (Molnar, 2020). Using this method, features that have a significant impact on Vs were verified. The 2D PDP is plotted in Fig. 8.
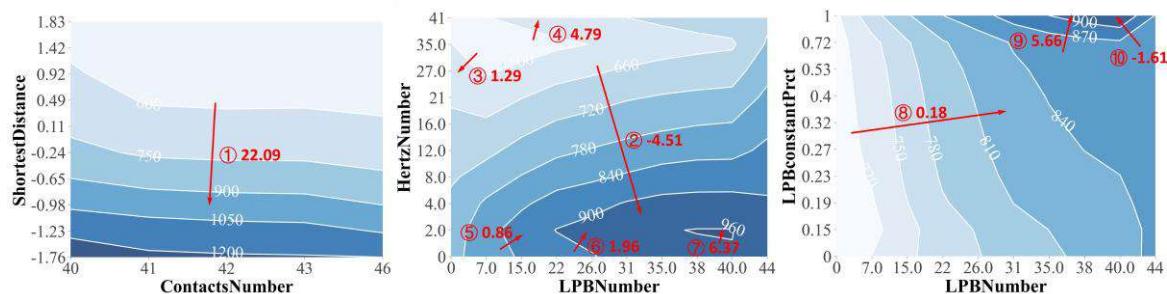
Fig. 8 2D PDP between bi-variate features, where red arrows represent the increasing direction of Vs and values represent the normalized contribution ratio of features ($F_{vertical}/F_{horizontal}$).

The 2D PDP between bi-variate features in Fig. 8 indicates the evolution of contributions of two features to Vs by quantifying the normal direction of the steep contour. In Fig. 8, though both *ContactsNumber* and *ShortestDistance* had a monotonic negative impact on predicted Vs, the interaction between the two features showed that the variation of *ContactsNumber* hardly affects the contribution of *ShortestDistance* to Vs, as the normalized contribution ratio was 22.09 (Arrow #1). *HertzNumber* and *LPBNumber* were negatively correlated and played similar importance generally (−4.51 in Arrow #2); this was also verified in Fig. 8.
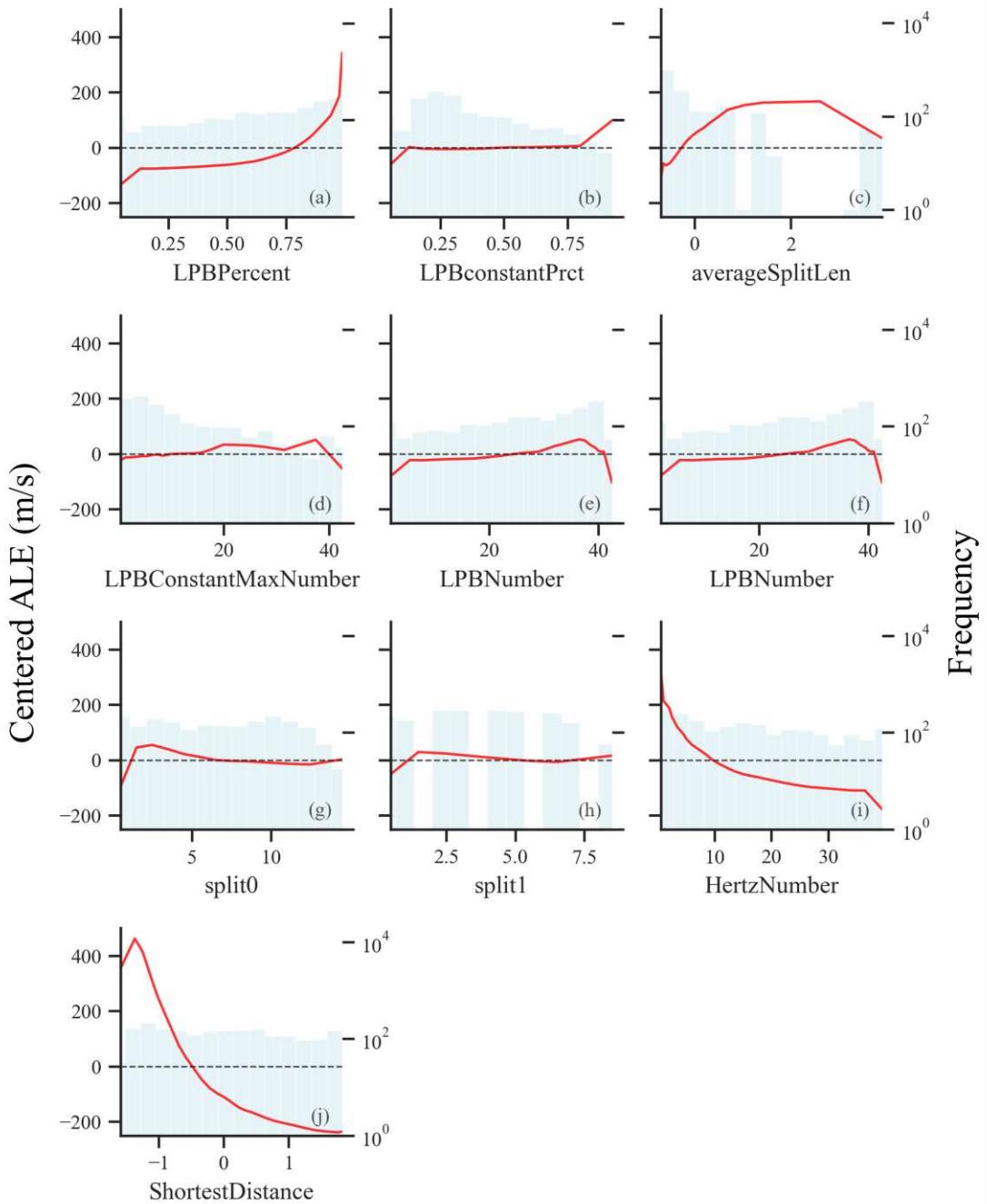
Yet, when *LPBNumber* was low and *HertzNumber* was high, reverse contributions occurred (Arrows #3, #4). While *HertzNumber* was low, the increase of *LPBNumber* contributed less and less to Vs with a ratio from 0.86 to 6.37 in Arrows #5, #6, and #7. *LPBNumber* and *LPBconstantPrct* were other feature pairs indicating the different evolutions. Vs kept increasing as *LPBNumber* and *LPBconstantPrct* increased, and *LPBNumber* contributed more when Vs was below 870 m/s (0.18 in Arrow #8). When Vs was high (>870 m/s) with

17

LPBNumber>31 and *LPBconstantPrct*>0.77, the increase of *LPBNumber* contributed little to the Vs increase and probably caused the opposite effect (5.66 and −1.61 in Arrows #9, #10). The result was interpreted as a high value of *LPBNumber,* indicating a long propagation distance; this resulted in a low Vs (Bate et al. 2021). Under these conditions, *LPBconstantPrct* had a greater influence on the increase in high Vs as the normalized contribution ratio was from 5.66 to -1.61.

4.3.2 ALE

Fig. 9 shows the ALE curves (Red line) of the trained AutoML model; the feature distribution is also presented. As mentioned above, ALE plots are a more unbiased alternative to PDP, especially when the features are not independent. *LPBPercent* had a significant and lasting positive impact on Vs with centered ALE from -134 to 343, while *HertzNumber* (centered ALE from 326 to -177) and *ShortestDistance* (centered ALE from 461 to -236) had the steepest negative cumulative effect. Analysis of *LPBconstantPrct* was also relevant as the feature's centered ALE echoed the role of cementation in different stages of Vs evolution (Sun et al. 2022). When the cementation content was low (*LPBconstantPrct* < 80%), the continuous contact was relatively dispersed; this resulted in the difficulty of searching for a faster propagation path for shear waves. With the percentage of continuous LPB contacts increased to over 80%, the centered ALE started rising from 6 to 97; this echoed that the high content of continuous LPB contacts greatly contributed to the Vs prediction (Sun et al. 2025).

18

Fig. 9 ALE curves of 10 features regarding the trained AutoML model.

4.3.3 LIME

Fig. 10 shows a local feature explanation of a set of instances ordered by cementation content in LIME. With the increase in cementation content, these features showed similar trends in contribution to Vs; *ShortestDistance* emerged as the most influential feature, with the negative contribution exceeding 300 m/s. Additionally, the *averageSplitLen* and *HertzNumber* of this sample also had a negative contribution greater than 100 m/s; this confirmed the importance of the number of different contact types and continuous particle cementation, respectively (Cheng et al. 2013; Yun & Evans 2011). Unlike other features, the lower value of *split1*, which represented the dispersion of cementation, positively contributed to Vs when Vs was lower

than 273 m/s; this indicates that too-dispersed cementation under low cementation content could only lead to a slight stiffness increase in uncemented sand samples, echoing the findings in Stage I in M. Sun et al. (2022).
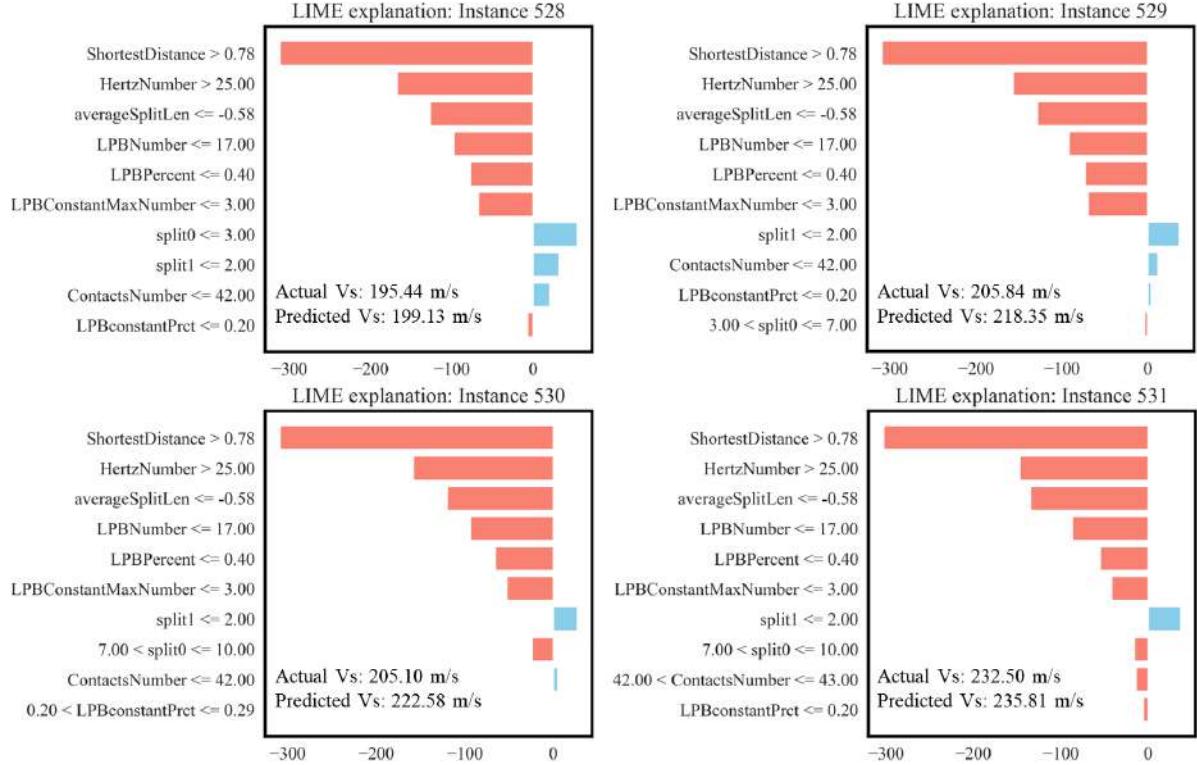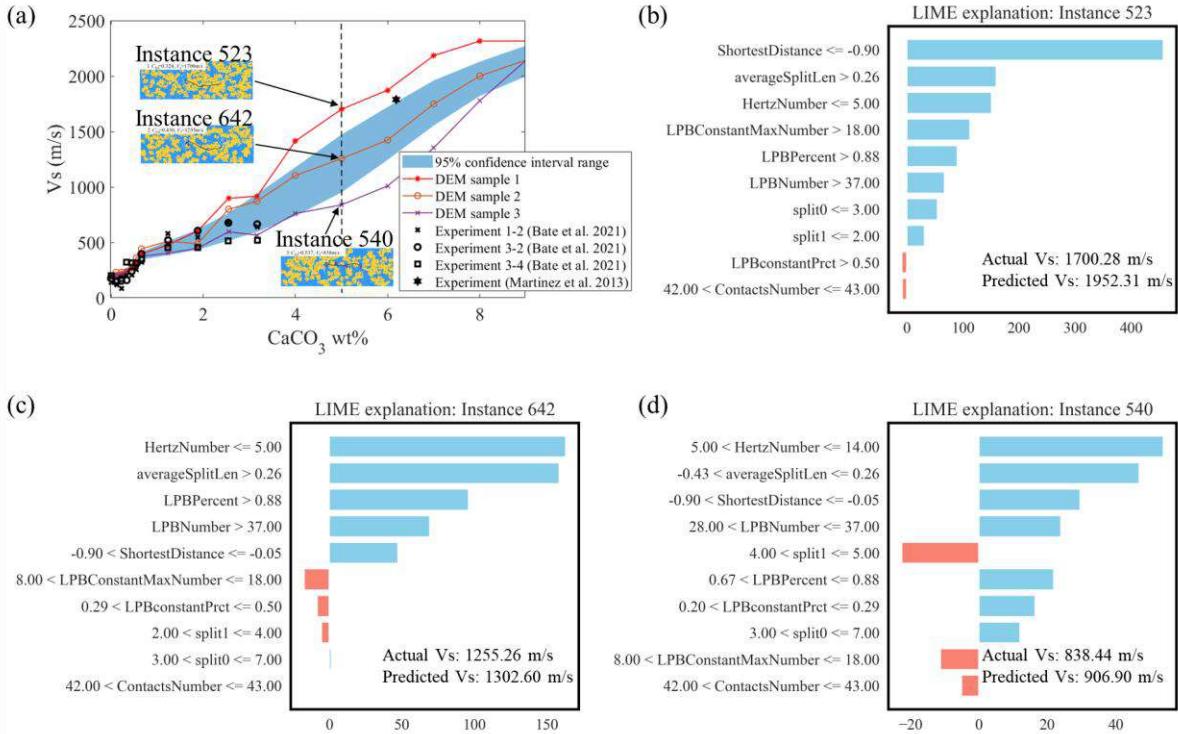
Fig. 10 The local feature explanation of 4 instances ordered by cementation content in LIME.

As shown in Fig. 11, Vs varied in the three instances, though the cementation content was the same (5%, Stage III). Instance No. 523 exhibited a notably higher Vs due to its superior *ShortestDistance* as described in the above interpretability analysis (Fig. 11b). On the other hand, the low Vs of the other two instances were attributed to the following reasons.

Fig. 11 Comparison of three instances under the same cementation content (5%, Stage III) by LIME. (a) The variation range of Vs adapted from M. Sun et al., 2025; LIME explanation of Instance 423 (b), 642 (c), and 540 (d).
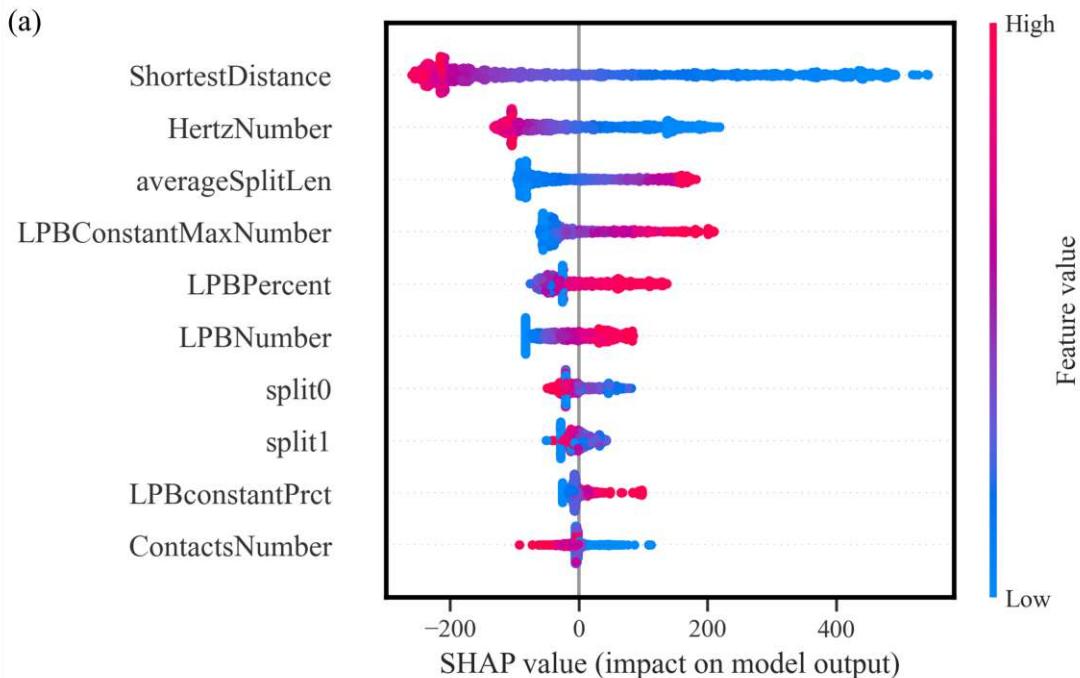
As shown in Fig. 11c, in Instance No. 642, the high *ShortestDistance* value was the primary cause of the low Vs; this reflected the primary importance of this feature (Sun et al. 2025). The high value of *averageSplitLen* and the low value of *HertzNumber* replaced the *ShortestDistance* as the dominant features. The result echoed the evidence showing that in the absence of direct linear distance as the shortest path, the shear wave preferred to choose a path with long-distance cementation and fewer Hertz contacts to speed up Vs (Sun et al. 2025).
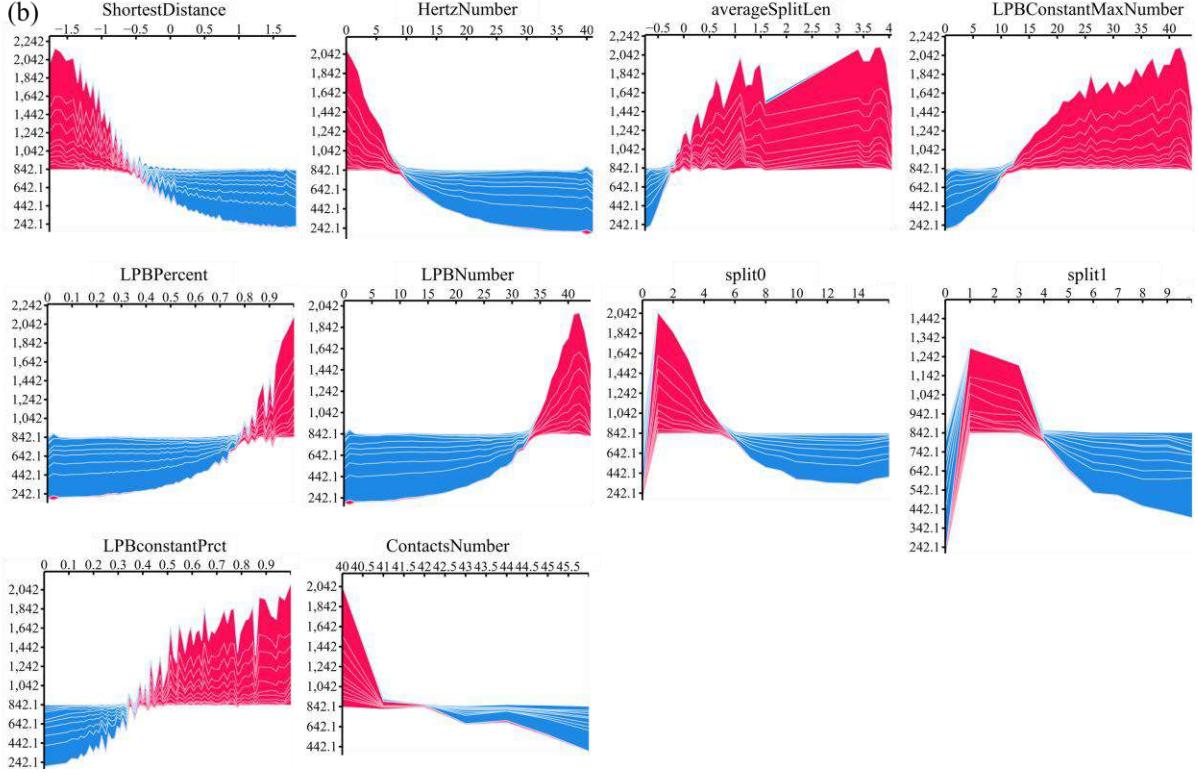
As shown in Fig. 11d, the *ShortestDistance* of Instance No. 540 still functions as a significant feature, but the incoherence of cementation in the path seriously reduces Vs. This is reflected in that *split1*, *LPBConstantMaxNumber,* and *split0,* which represent the incoherence of cementation, negatively affect Vs. Among them, the large negative effect of *split1* echoes that incoherent continuous cementation can significantly affect Vs (Sun et al. 2022).

To sum up, the comparison analysis of these three instances was consistent with prior studies (Sun et al. 2022; Sun et al. 2025). The cementation variability led to diverse Vs under the same cementation content, and the difference between the three instances was properly interpreted within the microscopic scope; this interpretation focused on the influence of cementation on the propagation path. This LIME method compensated for the lack of experience in the end-to-end black-box AutoML prediction; it provided a reasonable interpretability analysis of the Vs difference between the three instances.

21

4.3.4 SHAP

Fig. 12 shows the global importance of features by SHAP. The wide SHAP value range of *ShortestDistance* showed a significant impact on predicted Vs (Fig. 12a), consistent with the above analysis result. Low values of *ShortestDistance*, *HertzNumber*, *split0*, and *ContactsNumber* made a positive contribution generally, while the others (except *split1*) had the opposite effect. Fig. 12b specifically shows the predicted Vs evolution under different features. Although the predicted Vs was generally monotonically related to the increase of the selected feature, the evolution by *LPBConstantMaxNumber*, *LPBNumber*, *split0*, and *split1* had extreme values. The extreme values under *LPBConstantMaxNumber* = 42 and *LPBNumber* = 42 corresponded to the contact count in the shortest path. In addition, lower *split0* and *split1* values generally correlated with higher Vs (Sun et al. 2025). However, the extremal condition (*split0* = 1, *split1* = 1) specifically identified uncemented samples (i.e., *split0* = 0, *split1* = 0).
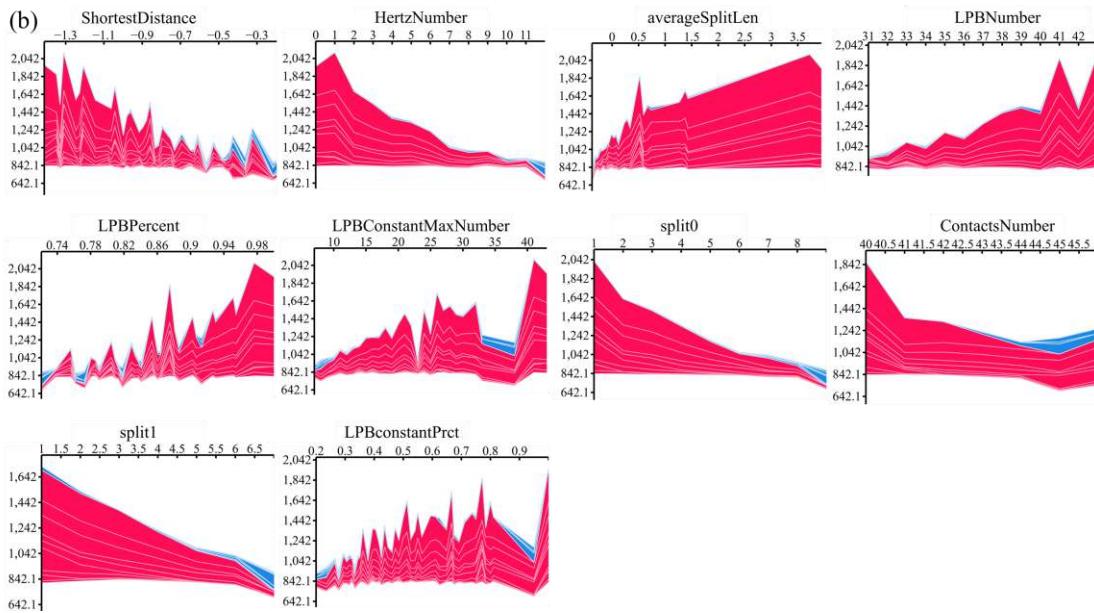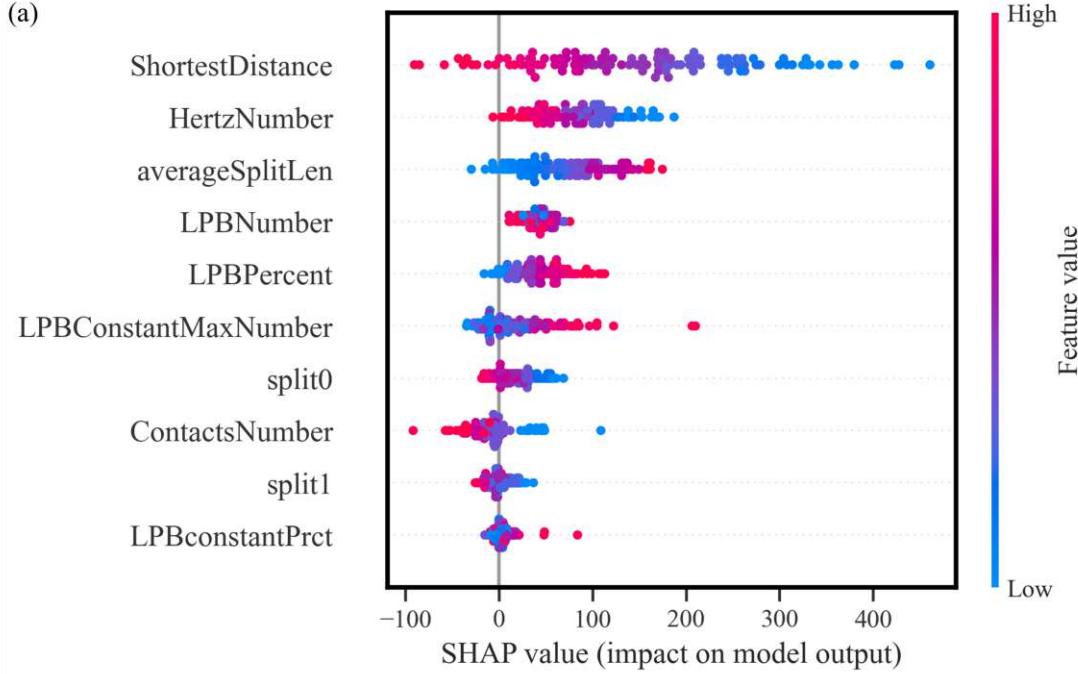
Fig. 12 Global summary of features in all samples. (a) Global feature beeswarm plot; (b) Force plot of predicted Vs versus feature values ranked by the selected feature.

Fig. 13 shows that SHAP also allowed for the contribution analysis of the above microscopic features on the variability of Vs in local instances, when the cementation content was 5%. In Fig. 13a, the high SHAP value range of *ShortestDistance* indicates that it was the main cause of Vs variability. On the other hand, *LPBNumber*, *split0*, *split1,* and *LPBconstantPrct* were not responsible for the variability of Vs, as these SHAP values had a small variation range. In Fig. 13b, the smooth evolution by *HertzNumber*, *averageSplitLen*, *LPBNumber*, *split0*, *ContactsNumber,* and *split1* indicates that the predicted Vs caused by changes in these features are continuous and gradual, which have a stable influence; the zigzag evolution by *ShortestDistance* and *LPBPercent* indicates these features would affect the prediction by interacting with other features; the more zigzag of *LPBConstantMaxNumber* and *LPBconstantPrct* indicate their complex relationship with predicted values when the cementation content was 5%.

23

Fig. 13 Local summary of features when cementation content was 5%. (a) feature beeswarm plot; (b) Force plot of predicted Vs versus feature values ranked by the selected feature

Fig. 14 compares the local analysis by SHAP between the above three instances presented in Fig. 11a. In Instance No. 523, *ShortestDistance* = -1.208 (+328.54m/s) contributed the most to predicted Vs (1952.31m/s), followed by *LPBConstantMaxNumber* = 43 (+210.11m/s); this was consistent with the explanation of that by LIME (Fig. 11b). In Instance No. 642, the positive contribution of *LPBConstantMaxNumber* = 17 (+1.38m/s) and *ShortestDistance* = -0.878 (+153.68m/s) to the predicted Vs was significantly reduced; this was the main reason for the low Vs (1302.06m/s). In Instance No. 540, the contributions of all features decreased significantly; this led to the lowest predicted Vs (906.90m/s) compared with the other two

instances. *ShortestDistance*, the most important feature in the global analysis, had essentially no positive contribution (+8.74m/s). Different from the LIME explanation, the difference of *LPBConstantMaxNumber* contributions varied significantly (+210.11m/s in Instance No. 523, +1.38m/s in Instance No. 642, and -19.01m/s in Instance No. 540); this increased the Vs gap between the three instances.
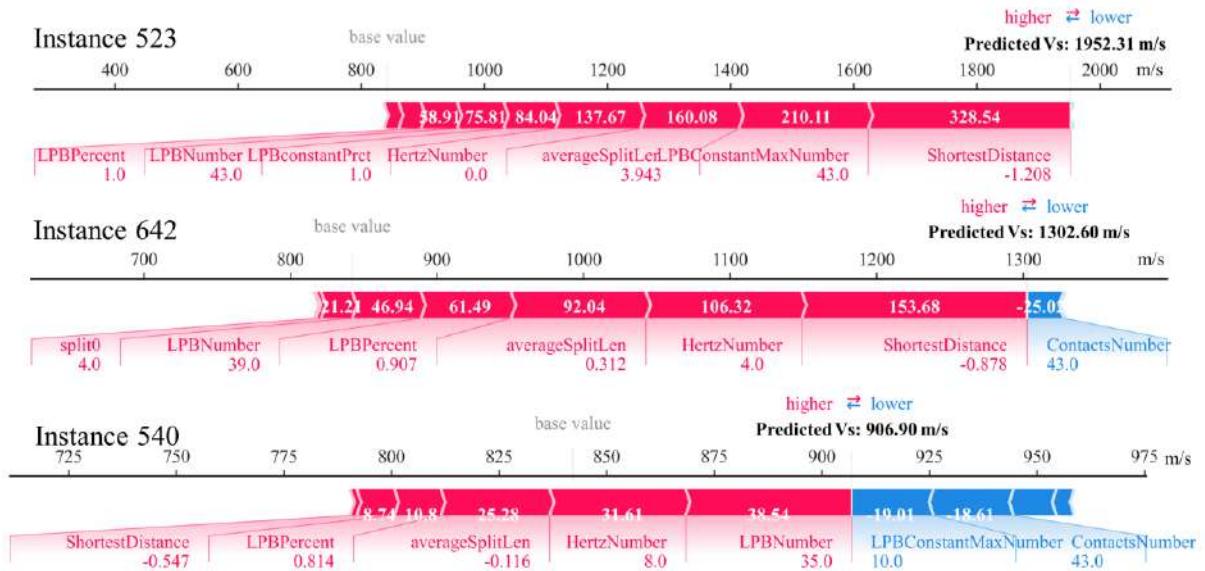

Fig. 14 Feature contribution comparison of 3 instances under the same cementation content (5%) by SHAP

4.4 Sensitivity analysis and feature ablation

Two analyses, namely sensitivity and ablation, were employed to further assess the feature importance. Sensitivity analysis indicates how perturbations in features affected the predicted Vs, and the ablation experiment reflects feature importance by removing features. These approaches provide quantitative evidence to support post-hoc interpretations.

The ICE curves in Fig. 7 for individual instances (light blue curves) show the overall predicted Vs changes and indicate the model sensitivity to feature perturbations. It is noted that predicted Vs is highly sensitive to *ShortestDistance* and *HertzNumber* in individual samples, with average changes of 750 m/s and 300 m/s, respectively. The evidence of high sensitivity and consistent patterns across samples strongly confirms their roles as key features influencing the effectiveness of cementation reinforcement.

A feature ablation analysis was also performed to further quantify the importance of each feature. This involved sequentially removing each feature, retraining the AutoML model, and recording the drop in the evaluation (ΔR^2). Consistent with the results of the interpretability analysis, the drop of *ShortestDistance* led to the greatest performance drop ($\Delta R^2 = -0.006$), followed by *HertzNumber* ($\Delta R^2 = -0.004$) and *LPBConstantMaxNumber* ($\Delta R^2 = -0.004$). In contrast, removing other features had little impact on performance ($|\Delta R^2| < 0.002$). The ablation confirms that the model relies heavily on *ShortestDistance* and *HertzNumber* to make accurate

25

predictions.

5 Discussion

5.1 AutoML-enhanced DEM simulation

The findings in this paper show that AutoML enhanced Vs prediction using the DEM simulated dataset. Due to the high costs of laboratory experiments and DEM simulations, coupled with the limitations of $C_{d,n}$-based methods in capturing sample instance variability, AutoML offered a necessary solution for efficient and accurate characterization. The proposed AutoML method achieved higher accuracy than $C_{d,n}$-based methods (Sun et al. 2025); it incorporated more microscopic features and nonlinear regression models to predict stiffness in static samples (Fig. 5). Plus, the automation of AutoML offered a convenient tool for geotechnical professionals and researchers. The open-source method can be readily extended to other relevant experiments and DEM simulations.

The evaluation of prediction uncertainty is valuable in guiding engineering practice. The proposed AutoML approach, with 90.16% of predictions falling within a 20% error interval, achieved a substantially higher percentage than $C_{d,n}$-Coord (78.50%) and $C_{d,n}$-MC (44.27%) methods. However, a small proportion of outliers (9.84%) reflects the inherent uncertainty associated with complex geotechnical systems. Therefore, in engineering practice, excessive outliers from the expected range should be treated with caution, and additional site-specific verification should be conducted if necessary.

5.2 Comparisons between the four interpretability methods

This paper used four model-agnostic methods to analyze AutoML's prediction results from different perspectives. Generally, the feature interpretations of the predicted Vs (Section 4.3) aligned with prior experimental observations, simulations, and shear wave propagation principles (Bate et al. 2021; Sun et al. 2025). Furthermore, the quantitative analysis provides researchers with critical information on key features for experiments and simulation designs.

Two global interpretability methods, PDP and ALE, presented similar feature importance rankings and complemented each other. While PDP reduced the persuasiveness of global interpretability due to the feature independence assumption, it uniquely revealed the interaction between two features. The alternative global method, ALE, addressed feature correlation and identified *ShortestDistance* and *HertzNumber* as the most significant features. These two features were consistent with the PDP results, and this consistency can be validated through DEM analysis (Sun et al. 2025).

Both LIME and SHAP, as local model-agnostic methods, focus on interpreting the differences between samples. Consistent with prior studies, *ShortestDistance* was the dominant factor in the differences between samples. However, cementation continuity in Stages I and III of cementation content also greatly affected shear wave propagation. Continuous cementation

26

allows for faster shear wave propagation than discontinuous cementation at the same *ShortestDistance* value in Stage III. The feature ranking in LIME interpretation (e.g., Fig. 10) varied for each instance. This variation indicates that LIME's local fidelity ensures a faithful local approximation but may not provide a good global approximation. SHAP addressed this limitation through dual global-local interpretability.

In summary, although each interpretability method offered a distinct yet limited perspective, their interpretations were not in conflict but were highly complementary instead. The advantage of ALE in handling correlated features complements the feature independence assumption of PDP. While LIME provided local fidelity, it was sensitive to kernel width and failed to provide a global interpretation of the models. SHAP addressed this limitation through combined global-local interpretability based on a game-theoretical framework. These complementary methods provided a robust predicted Vs interpretation, aligning with domain knowledge. Integrating these four methods allowed for a more trustworthy interpretation than any single method could offer.

5.3 Research significance

The AutoML prediction and interpretability analysis offer valuable feedback and guidance for experiments and DEM simulations (Zhou & Xue 2025). This analysis highlights features that warrant global attention in engineering practice and reinforcement research, such as *ShortestDistance* in Stages I and III. Specifically, these features have practical implications in the reinforcement process. For instance, the reinforcement process should maximize the total cementation content and achieve a continuous spatial distribution of cementation as suggested by the interpretation. Additionally, local interpretability methods allow for the analysis of sample-specific variability, which helps explain performance differences in individual field samples.

The significance of this research is summarized below:

1. This paper provides an end-to-end AutoML prediction approach that enhances computational efficiency in small-strain simulation, exceeding existing Vs quantification methods (e.g., $C_{d,n}$-based predictions) that only capture basic evolutionary trends. This efficient approach can be adopted in practical engineering to rapidly predict Vs without the need for costly simulations.
2. This paper quantifies the significance of each microscopic feature on the macroscopic Vs evolution by interpretability analysis. Four model-agnostic methods verify the consistency of the feature interpretations with expertise in experimental theory and DEM simulation. This quantification provides practitioners with prior features to monitor and control during the reinforcement process.

5.4 Limitations and future directions

Although the AutoML prediction and interpretability analysis contribute to the study of

27

cementation reinforcement mechanisms, the following limitations also exist:

1. Omission of the temporal dimension: The disordered time-series samples hinder the interpretation of the dynamic reinforcement process due to the loss of sample connections.
2. Deficiency in stage-wise analysis: Although global and instance-specific feature importance were analyzed, stage-based characteristics remained unclear; this limited the understanding of temporal evolutionary patterns in the reinforcement process (Sun et al. 2022).
3. Macroscopic parameter gap: The significance of macroscopic parameters for Vs was not investigated.

Future research directions can therefore be suggested based on the limitations above and general technological prospects:

1. Within the experiment and DEM simulation, the reinforcement process was a temporally injected solution that can be treated as a time-series dataset (Bate et al. 2021). These time-sequenced microscopic spatial features can be imported into recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) models to capture evolutionary patterns (Ghorbani et al. 2025; Wang et al. 2023; Rajabi et al. 2023; Hazbeh et al. 2024). Combining these temporal models with the current AutoML framework could create a multi-stage predictive model that accounts for both temporal and spatial variability. The proposed approach would provide a whole view of the entire time-dependent reinforcement process.
2. Stage-wise feature importance analysis should be emphasized as the significance of microscopic features differs between stages. Future work should include stage-by-stage sample analysis to better guide the reinforcement process. More multi-scale datasets should also be established to optimize the model.
3. Direct macro-to-micro prediction should be further studied. While this paper analyzed the effect of microscopic features on shear stiffness, advances in AI enable the direct utilization of easily accessible macroscopic features as inputs. It is promising to address the challenge of micro-parameter calibration and infer microscopic properties through multi-scale modeling approaches.

Emerging digital twin and large language model (LLM)-based technology can advance the research on cementation reinforcement (Luo et al. 2023; Zhao et al. 2025). Integrating numerical simulation and experimental sensing can reconstruct scenarios depicting material states across sections, and LLM-based agents can automate calibration and sample generation tasks. These technological applications hold significant promise.

6 Conclusions

The conventional discrete element method (DEM) simulations are computationally expensive for shear wave velocity (Vs) prediction in cementation analysis, while machine learning (ML)

is long challenged by interpretability in applying to DEM. This paper proposes an integrated approach that combines end-to-end AutoML prediction with interpretability analysis using DEM simulation, demonstrating its applicability in other relevant potential microscopic studies. The new approach resolves the computational complexity associated with Vs prediction and the dynamic small-strain analysis of microscopic feature effects in the DEM simulation.

The test dataset included 1,972 DEM samples with 10 key microscopic features. Compared to the $C_{d,n}$-based fitting approach, the adopted AutoML improved both the generalization analysis and prediction accuracy for Vs. Four interpretability methods—PDP, ALE, LIME, and SHAP—were used to provide both global and local microscopic interpretations; these methods address concerns about AutoML's black-box nature. These model-agnostic methods identified the most influential microscopic features (e.g., *ShortestDistance*) and clarified their functional relationships with the predicted Vs, thereby validating the impact of cementation on Vs.

The following conclusions are drawn from the findings:
1. The algorithmic integration of AutoML with DEM-generated datasets outperformed the traditional $C_{d,n}$-based fitting approach, and achieved a high R^2 of 0.952 with strong robustness.
2. The unified interpretability pipeline combining PDP, ALE, LIME, and SHAP methods verified that *ShortestDistance* and *HertzNumber* are two dominant microscopic features for predicting Vs, with cementation continuity (*averageSplitLen* and *LPBConstantMaxNumber*) affecting shear waves in Stages I and III of cementation.
3. The interpretable AutoML approach provided transparent and model-agnostic feedback that can be further applied to other potential microscopic studies.

The interpretable AutoML approach presented in this paper establishes a robust foundation for future research. Promising research directions can be extended, including incorporating a time-dependent reinforcement process, stage-wise feature importance analysis, and integrating additional macroscopic features.

Data Availability Statement
The DEM dataset in this study will be made publicly available through a GitHub repository.

Competing Interests

The authors declare that they have no competing interests.

CRediT Author Statement

Meng Sun – Data Curation, Methodology, Writing - Original Draft, Formal analysis, Investigation

Dong Liang – Formal analysis, Validation

Jiajia Wang – Visualization

Bate Bate – Resources

Fan Xue – Writing - Review & Editing, Supervision, Project administration, Funding acquisition, Conceptualization

References

Anemangely, M., Ramezanzadeh, A., Amiri, H. & Hoseinpour, S.-A. (2019, March). Machine learning technique for the prediction of shear wave velocity using petrophysical logs. *Journal of Petroleum Science and Engineering, 174*, 306–327. doi:10.1016/j.petrol.2018.11.032

Apley, D. W. & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 82*(4), 1059–1086. doi:10.1111/rssb.12377

Ashraf, M. S., Azahar, S. B. & Yusof, N. Z. (2017). Soil improvement using MICP and biopolymers: A review. *IOP Conference Series: Materials Science and Engineering, 226*, 012058. doi:10.1088/1757-899X/226/1/012058

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*, e0130140. doi:10.1371/journal.pone.0130140

Bang, S. S., Bang, S., Frutiger, S., Nehl, L. M. & Comes, B. L. (2009). Application of novel biological technique in dust suppression. *Transportation Research Board 88th Annual Meeting.* Washington DC. Retrieved from https://trid.trb.org/view/880867

Baratchi, M., Wang, C., Limmer, S., van Rijn, J. N., Hoos, H., Bäck, T. & Olhofer, M. (2024). Automated machine learning: past, present and future. *Artificial Intelligence Review, 57*(5), 122. doi:10.1007/s10462-024-10726-1

Barbudo, R., Ventura, S. & Romero, J. R. (2023). Eight years of AutoML: categorisation, review and trends. *Knowledge and Information Systems, 65*(12), 5097–5149. doi:10.1007/s10115-023-01935-1

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. doi:10.1016/j.inffus.2019.12.012

Bate, B., Cao, J., Zhang, C. & Hao, N. (2021). Spectral induced polarization study on enzyme induced carbonate precipitations: influences of size and content on stiffness of a fine sand. *Acta Geotechnica, 16*(3), 841–857. doi:10.1007/s11440-020-01059-8

Chang, I., Im, J. & Cho, G.-C. (2016). Introduction of microbial biopolymers in soil treatment for future environmentally-friendly and sustainable geotechnical engineering. *Sustainability, 8*(3), 251. doi:10.3390/su8030251

Cheng, L., Cord-Ruwisch, R. & Shahin, M. (2013). Cementation of sand soil by microbially induced calcite precipitation at various degrees of saturation. *Canadian Geotechnical Journal, 50*(1), 81-90. doi:10.1139/cgj-2012-0023

Dalloo, A. M. & Humaidi, A. J. (2024). Optimizing Machine Learning Models with Data-level Approximate Computing: The Role of Diverse Sampling, Precision Scaling, Quantization and Feature Selection Strategies. *Results in Engineering, 24*, 103451. doi:10.1016/j.rineng.2024.103451

De Bie, T., De Raedt, L., Hernández-Orallo, J., Hoos, H. H., Smyth, P. & Williams, C. K. (2022). Automating data science. *Communications of the ACM, 65*(3), 76–87. doi:10.1145/3495256

DeJong, J. T., Mortensen, B. M., Martinez, B. C. & Nelson, D. C. (2010). Bio-mediated soil improvement. *Ecological Engineering, 36*(2), 197–210. doi:10.1016/j.ecoleng.2008.12.029

Deka, P. (2019). *A primer on machine learning applications in civil engineering.* Boca Raton: CRC Press. doi:10.1201/9780429451423

Euldji, R., Bouamhdi, M., Rebhi, R., Bachene, M., Ikumapayi, O. M., Al-Dujaili, A. Q., Abdulkareem, A. I., Humaidi, A. J. & Menni, Y. (2023). Optimizing condition monitoring of ball bearings: An integrated approach using decision tree and extreme learning machine for effective decision-making. *Open Physics, 21*. doi:10.1515/phys-2022-0239

Feng, K., Montoya, B. M. & Evans, T. M. (2017). Discrete element method simulations of bio-cemented sands. *Computers and Geotechnics, 85*, 139–150. doi:10.1016/j.compgeo.2016.12.028

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M. & Hutter, F. (2022). Auto-Sklearn 2.0: Hands-free AutoML via meta-Learning. *Journal of Machine Learning Research, 23*(261), 1-61. doi:10.48550/arXiv.2007.04074

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M. & Hutter, F. (2019). Efficient and robust automated machine learning. In F. Hutter, J. Vanschoren & L. Kotthoff, *Automated Machine Learning: Methods, Systems, Challenges* (pp. 113-134). Springer. doi:10.1007/978-3-030-05318-5_6

Feurer, M., Springenberg, J. & Hutter, F. (2015). Initializing bayesian hyperparameter optimization via meta-learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*. doi:10.1609/aaai.v29i1.9354

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. doi:10.1214/aos/1013203451

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing, 21*(2), 137–146. doi:10.1007/s11222-009-9153-8

Geurts, P., Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42. doi:10.1007/s10994-006-6226-1

Ghorbani, H., Hazbeh, O., Rajabi, M., Tabasi, S., Lajmorak, S., Ahmadi Alvar, M. & Radwan, A. E. (2025, October). Prediction of formation shear wave velocity using a new hybrid deep

learning algorithm. *Physics and Chemistry of the Earth, Parts A/B/C, 140*, 104063. doi:10.1016/j.pce.2025.104063

Han, D.-h., Nur, A. & Morgan, D. (1986, November). Effects of porosity and clay content on wave velocities in sandstones. *GEOPHYSICS, 51*, 2093–2107. doi:10.1190/1.1442062

Hazbeh, O., Rajabi, M., Tabasi, S., Lajmorak, S., Ghorbani, H., Radwan, A. E., Alvar, M. A. & Molaei, O. (2024, April). Determination and investigation of shear wave velocity based on one deep/machine learning technique. *Alexandria Engineering Journal, 92*, 358–369. doi:10.1016/j.aej.2024.03.007

Hu, J. & Wang, J. (2024). A data extension framework of seismic-induced gravelly soil liquefaction based on semi-supervised methods. *Advanced Engineering Informatics, 59*, 102295. doi:10.1016/j.aei.2023.102295

Huang, S., Huang, M. & Lyu, Y. (2019). A novel approach for sand liquefaction prediction via local mean-based pseudo nearest neighbor algorithm and its engineering application. *Advanced Engineering Informatics, 41*, 100918. doi:10.1016/j.aei.2019.04.008

Jiang, M.-J. (2019). New paradigm for modern soil mechanics: Geomechanics from micro to macro. *Chinese Journal of Geotechnical Engineering, 41*(2), 195–254. doi:10.11779/CJGE201902001

Karol, R. H. (2003). *Chemical Grouting and Soil Stabilization.* Boca Raton, FL, USA: CRC Press. doi:10.1201/9780203911815

Kempfert, H.-G. & Gebreselassie, B. (2006). *Excavations and foundations in soft soils.* Berlin/Heidelberg: Springer. doi:10.1007/3-540-32895-5

Kim, B., Khanna, R. & Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2288–2296). Barcelona: Curran Associates. doi:10.5555/3157096.3157352

Li, J., Wang, H., Luo, H., Jiang, X. & Li, E. (2023). A ranking prediction strategy assisted automatic model selection method. *Advanced Engineering Informatics, 57*, 102068. doi:10.1016/j.aei.2023.102068

Liang, D. & Xue, F. (2023). Integrating automated machine learning and interpretability analysis in architecture, engineering and construction industry: A case of identifying failure modes of reinforced concrete shear walls. *Computers in Industry, 147*, 103883. doi:10.1016/j.compind.2023.103883

Lin, H., Suleiman, M. T., Brown, D. G. & Kavazanjian, E. (2016). Mechanical behavior of sands treated by microbially induced carbonate precipitation. *Journal of Geotechnical and Geoenvironmental Engineering, 142*(2), 04015066. doi:10.1061/(ASCE)GT.1943-5606.0001383

Liu, X., Yang, J., Zou, D., Li, Z., Chen, Y. & Cao, X. (2024). Utilizing DEM and interpretable ML algorithms to examine particle size distribution's role in small-strain shear modulus of gap-graded granular mixtures. *Construction and Building Materials, 428*, 136232. doi:10.1016/j.conbuildmat.2024.136232

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768–4777). Red Hook, NY, United States: Curran Associates Inc. doi:10.5555/3295222.3295230

Luo, X., Li, H. & Lee, S. (2023). Bridging the gap: Neuro-Symbolic Computing for advanced AI applications in construction. *Frontiers of Engineering Management, 10*(4), 727–735. doi:10.1007/s42524-023-0266-0

Makarian, E., Mirhashemi, M., Elyasi, A., Mansourian, D., Falahat, R., Radwan, A. E., El-Aal, A., Fan, C. & Li, H. (2023, November). A novel directional-oriented method for predicting shear wave velocity through empirical rock physics relationship using geostatistics analysis. *Scientific Reports, 13*, 19872. doi:10.1038/s41598-023-47016-9

Marcílio, W. E. & Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images*, (pp. 340–347). Porto de Galinhas. doi:10.1109/SIBGRAPI51738.2020.00053

Martinez, B. C. & DeJong, J. T. (2009). Bio-mediated soil improvement: load transfer mechanisms at the micro-and macro-scales. In *Advances in Ground Improvement* (pp. 242–251). doi:10.1061/41025(338)26

Martinez, B. C., Dejong, J. T., Ginn, T. R., Montoya, B. M., Barkouki, T. H., Hunt, C., Tanyu, B. & Major, D. (2013). Experimental optimization of microbial-induced carbonate precipitation for soil improvement. *Journal of Geotechnical and Geoenvironmental Engineering, 139*(4), 587–598. doi:10.1061/(ASCE)GT.1943-5606.0000787

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*(4), 1–38. doi:10.1016/j.artint.2018.07.007

Molnar, C. (2020). *Interpretable machine learning.* Retrieved from https://christophm.github.io/interpretable-ml-book/

Niu, Q. & Zhang, C. (2018). Joint inversion of NMR and SIP data to estimate pore size distribution of geomaterials. *Geophysical Journal International, 212*(3), 1791–1805. doi:10.1093/gji/ggx501

Niu, W., Guo, B., Li, K., Ren, Z., Zheng, Y., Liu, J., Lin, H. & Men, X. (2024). Cementitous material based stabilization of soft soils by stabilizer: Feasibility and durabiliy assessment. *Construction and Building Materials, 425*, 136046. doi:10.1016/j.conbuildmat.2024.136046

Rajabi, M., Hazbeh, O., Davoodi, S., Wood, D. A., Tehrani, P. S., Ghorbani, H., Mehrad, M., Mohamadian, N., Rukavishnikov, V. S. & Radwan, A. E. (2023, January). Predicting shear wave velocity from conventional well logs with deep and hybrid machine learning algorithms. *Journal of Petroleum Exploration and Production Technology, 13*, 19–42. doi:10.1007/s13202-022-01531-z

Rasmussen, C. E., Bousquet, O., von Luxburg, U. & Rätsch, G. (2004). Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning* (pp. 63–71). Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-28650-9_4

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *ArXiv Preprint*. doi:10.48550/arXiv.1606.05386

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016b). "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* (pp. 1135-1144). Association for Computing Machinery. doi:10.1145/2939672.2939778

Ribeiro, M. T., Singh, S. & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32.* doi:10.1609/aaai.v32i1.11491

Sadd, M. H., Adhikari, G. & Cardoso, F. (2000, April). DEM simulation of wave propagation in granular materials. *Powder Technology, 109*, 222–233. doi:10.1016/S0032-5910(99)00238-7

Saka, A., Oyedele, L. O., Àkànbí, L., Ganiyu, S., Chan, D. W. & Bello, S. (2023). Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities. *Advanced Engineering Informatics, 55*, 101869. doi:10.1016/j.aei.2022.101869

Salifu, E., MacLachlan, E., Iyer, K. R., Knapp, C. W. & Tarantino, A. (2016). Application of microbially induced calcite precipitation in erosion mitigation and stabilisation of sandy soil foreshore slopes: A preliminary investigation. *Engineering Geology, 201*, 96–105. doi:10.1016/j.enggeo.2015.12.027

Santamarina, J. C., Klein, A. & Fam, M. A. (2001). Soils and waves: Particulate materials behavior, characterization and process monitoring. *Journal of Soils and Sediments, 1*(2), 130–130. doi:10.1007/BF02987719

Singh, P., Kapur, P. K., Singh, G. & Panwar, S. (2021). Data mining techniques and its application in civil engineering–A Review. In *Advances in Interdisciplinary Research in Engineering and Business Management* (pp. 175–183). Singapore: Springer Nature. doi:10.1007/978-981-16-0037-1_15

Sun, B., Cui, W., Liu, G., Zhou, B. & Zhao, W. (2023). A hybrid strategy of AutoML and SHAP for automated and explainable concrete strength prediction. *Case Studies in Construction Materials, 19*, e02405. doi:10.1016/j.cscm.2023.e02405

Sun, M., Cao, J., Cao, J., Zhang, S., Chen, Y. & Bate, B. (2022). Discrete element modeling of shear wave propagation in carbonate precipitate-cemented particles. *Acta Geotechnica, 17*(7), 2633-2649. doi:10.1007/s11440-022-01456-1

Sun, M., Liu, P., Chen, Y., Bate, B. & Xue, F. (2025). Sand stiffness variability induced by stochastic distributions of calcite precipitates: a Monte Carlo-DEM study. *Acta Geotechnica, 20*, 1363–1377. doi:10.1007/s11440-025-02539-5

Sun, M., Zhao, R. & Xue, F. (2024). The Opportunities and Challenges of Multimodal GenAI in the Construction Industry: A Brief Review. *Proceedings of the 29th International Symposium on Advancement of Construction Management and Real Estate.*

Varshney, K. (2016). Engineering safety in machine learning. *2016 Information Theory and Applications Workshop (ITA)*, (pp. 1-5). doi:10.1109/ITA.2016.7888195

Wang, L., Xiao, T., Liu, S., Zhang, W., Yang, B. & Chen, L. (2023, November). Quantification of model uncertainty and variability for landslide displacement prediction based on Monte

Carlo simulation. *Gondwana Research, 123*, 27–40. doi:10.1016/j.gr.2023.03.006

Xu, X., Qi, C., Aretxabaleta, X. M., Ma, C., Spagnoli, D. & Manzano, H. (2024). The initial stages of cement hydration at the molecular level. *Nature Communications, 15*(1), 2731. doi:10.1038/s41467-024-46962-w

Yang, P., O'Donnell, S., Hamdan, N., Kavazanjian, E. & Neithalath, N. (2017). 3D DEM simulations of drained triaxial compression of sand strengthened using microbially induced carbonate precipitation. *International Journal of Geomechanics, 17*(6), 04016143. doi:10.1061/(ASCE)GM.1943-5622.0000848

Yun, T. S. & Evans, T. M. (2011). Evolution of at-rest lateral stress for cemented sands: experimental and numerical investigation. *Granular Matter, 13*(5), 671–683. doi:10.1007/s10035-011-0279-y

Zhang, K., Min, Z., Hao, X., Henning, H. & Huang, W. (2025). Enhancing understanding of asphalt mixture dynamic modulus prediction through interpretable machine learning method. *Advanced Engineering Informatics, 65*, 103111. doi:10.1016/j.aei.2025.103111

Zhang, L., Wang, Y., Fu, X., Song, X. & Lin, P. (2024). Geological risk prediction under uncertainty in tunnel excavation using online learning and hidden Markov model. *Frontiers of Engineering Management*. doi:10.1007/s42524-024-0082-1

Zhang, X., Wang, L., Helwig, J., Luo, Y., Fu, C., Xie, Y., Liu, M., Lin, Y., Xu, Z., Yan, K., Adams, K., Weiler, M., Li, X., Fu, T., Wang, Y., Strasser, A., Yu, H., Xie, Y., Fu, X., Xu, S., Liu, Y., Du, Y., Saxton, A., Ling, H., Lawrence, H., Stärk, H., Gui, S., Edwards, C., Gao, N., Ladera, A., Wu, T., Hofgard, E. F., Tehrani, A. M., Wang, R., Daigavane, A., Bohde, M., . . . Ji, S. (2023). Artificial intelligence for science in quantum, atomistic, and continuum systems. *ArXiv Preprint*. doi:10.48550/arXiv.2307.08423

Zhao, Q. & Hastie, T. (2021). Causal interpretations of black-Box models. *Journal of Business & Economic Statistics, 39*(1), 272–281. doi:10.1080/07350015.2019.1624293

Zhao, X., Chen, P. & Tang, L. C. (2025). Condition-based maintenance via Markov decision processes: A review. *Frontiers of Engineering Management*. doi:10.1007/s42524-024-4130-7

Zhao, Y., Chen, X., Wen, T., Wang, P. & Li, W. (2022). Experimental investigations of hydraulic and mechanical properties of granite residual soil improved with cement addition. *Construction and Building Materials, 318*, 126016. doi:10.1016/j.conbuildmat.2021.126016

Zhou, Q. & Xue, F. (2025). Automatic information gain-guided convergence for refining building design parameters: Enhancing effectiveness and interpretability in simulation-based optimization. *Building and Environment, 275*, 112788. doi:10.1016/j.buildenv.2025.112788

Appendix A: Pseudocode of integrating automated machine learning and interpretability analysis

Algorithm S1 Pseudocode for integrating automated machine learning and interpretability analysis

Input: DEM simulation features ($\mathbf{F_i}$), Number of samples N

Output: Trained ensemble model (M), Vs predictions ($\widehat{Vs}$), Feature interpretation

 # Phase 1: Data Generation

for i=1 **to** N **do**

    Run DEM simulation with cementation parameters $CaCO_3$%, X_{seed}

    Extract microscopic feature vectors $\mathbf{F_i}$ and measured Vs

end for

Construct dataset D={($\mathbf{F_i}$, Vs)}$_{i=1}^{N}$

Standardize each feature in D

 # Phase 2: AutoML Modeling & Validation

Split D into k folds for cross-validation

for each fold (D_{train}, D_{test}) **do**

    Initialize Auto-sklearn regressor

    Fit model on D_{train} // Automated pipeline incl. feature preprocessing, model selection, and hyperparameter tuning

    Predict $\widehat{Vs}$ on D_{test})

    Evaluate performance (R^2)

end for

Construct the final ensemble model M from all trained models

 # Phase 3: Interpretability Analysis

for each interpretability method {PDP, ALE, LIME, SHAP} **do**

    Conduct interpretation for model M using dataset D

end for

return M, $\widehat{Vs}$, feature interpretation

Appendix B: Sample data

The sample data distribution matrix of these 1,972 samples is as follows (Sun et al., 2025):

$$
\begin{pmatrix}
S_{1\text{-}1} & S_{2\text{-}1} & \cdots & S_{116\text{-}1} \\
S_{1\text{-}2} & S_{2\text{-}2} & \cdots & S_{116\text{-}2} \\
S_{1\text{-}3} & S_{2\text{-}3} & \cdots & S_{116\text{-}3} \\
\vdots & \vdots & \ddots & \vdots \\
S_{1\text{-}17} & S_{2\text{-}17} & \cdots & S_{116\text{-}17}
\end{pmatrix}
=
\begin{pmatrix}
F(X_{\text{seed}_1},\text{CaCO}_3\%_1) & F(X_{\text{seed}_2},\text{CaCO}_3\%_1) & \cdots & F(X_{\text{seed}_{116}},\text{CaCO}_3\%_1) \\
F(X_{\text{seed}_1},\text{CaCO}_3\%_2) & F(X_{\text{seed}_2},\text{CaCO}_3\%_2) & \cdots & F(X_{\text{seed}_{116}},\text{CaCO}_3\%_2) \\
F(X_{\text{seed}_1},\text{CaCO}_3\%_3) & F(X_{\text{seed}_2},\text{CaCO}_3\%_3) & \cdots & F(X_{\text{seed}_{116}},\text{CaCO}_3\%_3) \\
\vdots & \vdots & \ddots & \vdots \\
F(X_{\text{seed}_1},\text{CaCO}_3\%_{17}) & F(X_{\text{seed}_2},\text{CaCO}_3\%_{17}) & \cdots & F(X_{\text{seed}_{116}},\text{CaCO}_3\%_{17})
\end{pmatrix}
\sim
\begin{pmatrix}
V_{S_{1\text{-}1}} & V_{S_{2\text{-}1}} & \cdots & V_{S_{116\text{-}1}} \\
V_{S_{1\text{-}2}} & V_{S_{2\text{-}2}} & \cdots & V_{S_{116\text{-}2}} \\
V_{S_{1\text{-}3}} & V_{S_{2\text{-}3}} & \cdots & V_{S_{116\text{-}3}} \\
\vdots & \vdots & \ddots & \vdots \\
V_{S_{1\text{-}17}} & V_{S_{2\text{-}17}} & \cdots & V_{S_{116\text{-}17}}
\end{pmatrix}.
$$

Matrix 1: sample number Matrix 2: model expression Matrix 3: the estimator Vs

In Matrix 1, a total of 17 × 116 samples were generated, where the number of columns (116) represents the sample group, and the number of rows (17) represents the sample numbering in each group. Matrix 2 demonstrates mathematically that each sample was determined by two variables: X_{seed}, the random-seed variable in the DEM simulation (representing the spatial distribution of cementation), ranging from 1 to 116; and $\text{CaCO}_3\%$, the calcite content in the DEM simulation (**Error! Not a valid bookmark self-reference.**). Through modeling each sample in Matrix 2, the Vs estimator matrix can be obtained, as shown in Matrix 3. All 1,972 DEM samples conformed to physical laws, therefore ensuring the representativeness in the context of shear wave velocity prediction.

Table S1 Calcite contents of 116 groups

m^*	$\text{CaCO}_3\%$	m^*	$\text{CaCO}_3\%$
1	0	10	2.55
2	0.11	11	3.17
3	0.23	12	4
4	0.34	13	5
5	0.45	14	6
6	0.55	15	7
7	0.66	16	8
8	1.24	17	9
9	1.88		

The m refers to the $\text{CaCO}_3\%$ numbering.

## Appendix C: Distributions of the ten microscopic features
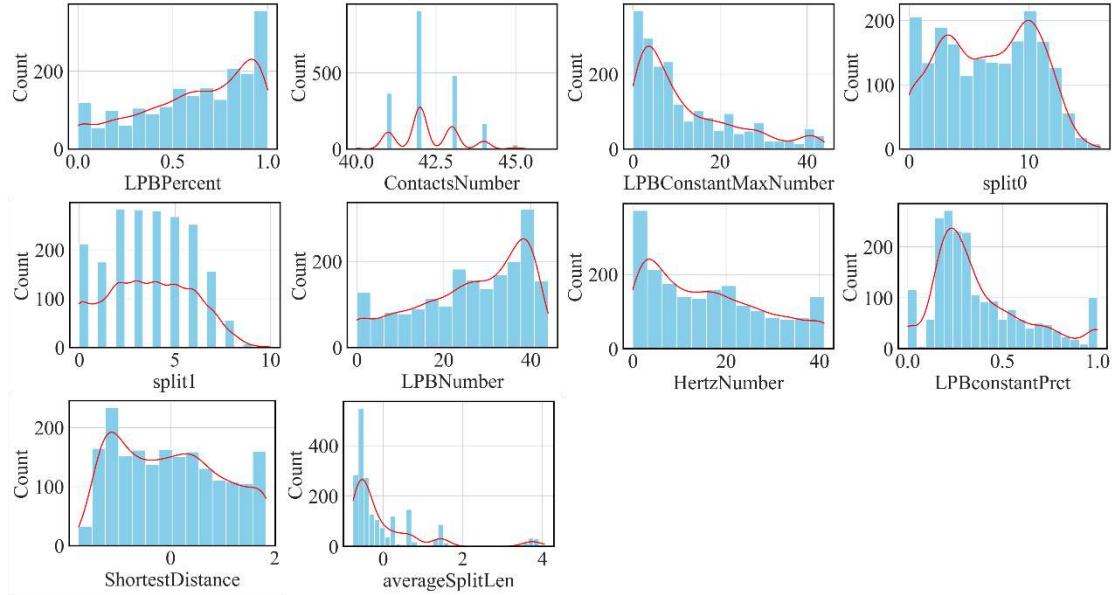
Fig. S1 Distributions of the ten microscopic features.

Appendix D: Summary of the optimal hyperparameter values in the ensemble model

Table S2 Summary of the optimal hyperparameter values in the ensemble model

Model	Weight	Parameter	Value
Gaussian_process	0.34	alpha	0.0560
		Kernel	RBF(length_scale = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
		n_restarts_optimizer	10
		normalize_y	True
		random_state	1
Gaussian_process	0.28	alpha	0.429
		Kernel	RBF(length_scale = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
		n_restarts_optimizer	10
		normalize_y	True
		random_state	1
K_nearest_neighbors	0.34	n_neighbors	20
		weights	distance
Adaboost	0.34	base_estimator	DecisionTreeRegressor(max_depth = 10)
		learning_rate	0.678
		n_estimators	175
		random_state	1