Journal of Engineering, Project, and Production Management
2026-16(7), 2026-0101

Numerical Simulation and Risk Assessment for Geological Engineering Stability in Construction Projects Using Machine Learning

Haicheng Pan[1] and Gong Faming[2]
[1] Ph.D. candidate, China University of Petroleum (East China), China
[2] Senior Industrial Internet Engineer, College of Computer Science and Technology, Qingdao Institute of Software, China University of Petroleum, Shandong Province, 266580, E-mail: uqpw9566@outlook.com (corresponding author).

Engineering Management
Received June 15, 2026; revised August 4, 2026; accepted August 10, 2026
Available online November 5, 2026

Abstract: Geological engineering stability is a primary determinant of safety, schedule, and cost performance in large construction projects, yet conventional limit-equilibrium and finite-element procedures struggle to quantify risk under parameter uncertainty and time-varying loading. This study proposes a hybrid Numerical-Simulation and Machine-Learning (NS-ML) framework that couples physics-based Finite-Element Modeling (FEM) of slopes, tunnels, and foundations with a stacked ensemble learner for quantitative stability classification and probabilistic risk assessment. A balanced dataset of 4,812 instances is assembled by augmenting 1,684 published field case histories with 3,128 calibrated FEM realizations spanning slope, tunnel, and shallow-foundation geometries. Six geomechanical descriptors feed the ensemble: unit weight, cohesion, friction angle, slope/overburden ratio, pore-pressure ratio and seismic coefficient. The base learners are Extreme Gradient Boosting (XGBoost), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and a Bidirectional Long Short-Term Memory (BiLSTM) network. These base learners are combined into a stacked ensemble using a logistic regression meta-learner. The proposed NS-ML model attains 95.7% classification accuracy, an F1-score of 0.954, and an AUC of 0.978 on a held-out test set, outperforming the strongest single learner (XGBoost, 92.4%) by 3.3 percentage points. SHAP-based interpretation identifies the friction angle and pore-pressure ratio as dominant risk drivers, while an ablation study quantifies the marginal value of each base learner and the FEM-augmented data partition. A Monte Carlo risk assessment module then converts the ensemble's probabilistic output into project-level expected loss, providing practical decision support for construction risk management.

Keywords: Geological engineering stability; numerical simulation; machine learning; ensemble learning; construction risk assessment; finite-element method.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-0101

1. Introduction

Geological engineering stability, encompassing slope failures, tunnel convergence, foundation settlement, and excavation-induced ground movement, remains one of the most consequential sources of risk in modern construction projects (Pradhan et al., 2023; Lopez and Garcia, 2024). Major projects routinely report cost overruns and schedule disruptions traceable to unforeseen geological behavior, and a non-trivial fraction of fatal construction incidents originate at the soil–structure interface (Zhang et al., 2022b,c). Reliable, quantitative prediction of stability under realistic uncertainty is therefore both an engineering and a project-management imperative.

Traditional risk assessment relies on limit-equilibrium analysis, the Finite-Element Method (FEM), and the Discrete-Element Method (DEM) (Li et al., 2022; Sanchez et al., 2025). These physics-based tools are mature, but their predictive fidelity is highly sensitive to input parameters that are seldom fully observed: spatial variability in strength, anisotropic seepage, joint orientation, and dynamic loading (Taherdangkoo et al., 2025; Bui et al., 2024). Calibration is laborious, and Monte-Carlo enrichment of FEM rapidly becomes computationally prohibitive at the scale of an entire construction program (Li et al., 2022; Kim et al., 2024).

Machine Learning (ML) addresses several of these limitations. Tree-based ensembles such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost), as well as deep learners such as Convolutional Neural Networks (CNN) and Long

Short-Term Memory (LSTM) networks, have been used to predict slope factor of safety, tunnel convergence, and settlement with high accuracy (Huang et al., 2023; Nasiri et al., 2026; Wang et al., 2023). Comparative reviews report that data-driven models can reach 85–94% classification accuracy on slope datasets, often outperforming individual numerical solvers (Phoon and Zhang, 2023; Sahoo et al., 2025). However, ML models trained purely on field case histories suffer from class imbalance, geographic bias, and limited generalization to construction settings outside the training distribution (Ahmad et al., 2024; Fatty et al., 2023).

Related advances in stacking-ensemble optimization and hybrid AI-driven sensing further motivate the use of ensemble surrogates for physically grounded prediction tasks (Rampriya et al., 2025; Vidhya et al., 2026; Elsayed et al., 2025; Mahalakshmi et al., 2025).

Recent work therefore emphasizes hybrid numerical-ML pipelines, in which FEM, DEM, or analytical solvers are used to synthesize additional, physically consistent training data, and ML is used to learn a fast surrogate or risk classifier on top of that data (Kim et al., 2024; Sanchez et al., 2025; Khan et al., 2025). Despite their promise, three gaps remain. First, most hybrid studies focus on a single geometry (e.g., slopes only), thereby limiting transferability to multi-asset construction projects that simultaneously involve slopes, tunnels, and foundations (Ma et al., 2024; Li et al., 2024). Second, ensemble combinations that combine tree-based learners with sequential deep learners remain underexplored in the geotechnical literature (Kumar et al., 2025; Liu et al., 2023). Third, project-level risk-assessment outputs, expected loss, probability of exceedance, and decision thresholds, are seldom reported, even though they are exactly what construction managers need (Shao et al., 2025; Phoon and Zhang, 2023).

1.1. Importance of Topic

Geological hazards directly govern the safety of workers and the viability of capital projects. In China alone, large open-pit mines report a 15% probability of landslide-class events (Wang et al., 2023). Tunneling projects routinely encounter unexpected fault zones, water intrusion, and mixed ground conditions that compromise schedules and budgets (Kim et al., 2024; Bai et al., 2022). Quantitative, data-driven stability assessment can transform reactive safety management into proactive risk control (Pradhan et al., 2023; Lopez and Garcia, 2024).

1.2. Research Gap

There is need for: (i) a unified framework that simultaneously handles slope, tunnel, and foundation stability. (ii) A principled coupling of FEM simulation with stacked ML ensembles to overcome data scarcity and class imbalance. (iii) An explicit translation of model outputs into project-level risk metrics suitable for construction decision-making.

1.3. Research Questions and Objectives

This research develops three research questions.

RQ1: How does a stacked ML ensemble trained on FEM-augmented data perform relative to single-learner baselines for multi-asset geological stability classification?

RQ2: What geomechanical descriptors and base learners contribute most to predictive performance?

RQ3: How do probabilistic stability outputs translate into project-level expected loss and decision thresholds?

Objectives: The objectives of this study are to (1) design a hybrid NS-ML framework that integrates FEM with a stacked ensemble. (2) Construct a balanced multi-asset dataset by augmenting field cases with FEM realizations. (3) Train and evaluate the proposed model against five competitive baselines. (4) Conduct an ablation study quantifying the contribution of each component. (5) Demonstrate how the framework supports construction-project risk assessment through Monte-Carlo expected-loss analysis.

Contributions: The principal contributions are: (i) a unified NS-ML architecture for multi-asset geological stability. (ii) An FEM-based data-augmentation protocol that mitigates class imbalance and geographic bias. (iii) A stacked ensemble that mixes tree-based and sequential deep learners and that achieves state-of-the-art accuracy on the assembled benchmark; and (iv) A Monte-Carlo risk-assessment layer that links model output to construction-management decisions.

Novelty: Unlike prior single-asset or purely data-driven studies, the novelty of this work lies in a single, tightly integrated pipeline. This pipeline (a) unifies slope, tunnel, and foundation stability under one label space, (b) couples calibrated FEM augmentation with a heterogeneous tree-plus-sequential stacked ensemble, and (c) closes the loop from a calibrated instability probability to a project-level expected-loss decision. This end-to-end organization, rather than any single component, constitutes the original contribution and structures the balance of the paper.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed methodology, including the theoretical background, the experimental setup, and the algorithms. Section 4 presents experiments and ablation studies. Section 5 discusses findings, limitations, and significance, and concludes the paper with an outline of future research.

2. Related Work

2.1. Numerical Simulation in Geotechnical Stability

The finite-element method has been the dominant tool for geotechnical stability analysis since the 1960s, supporting stress, displacement, and failure analyses for slopes, tunnels, and dams (Li et al., 2022). Recent advances integrate FEM with grey-system theory and uncertainty quantification to handle incomplete parameter information (Li et al., 2022), and with physics-

informed neural networks (PINNs) to solve coupled hydro-mechanical equations more efficiently (Sanchez et al., 2025; Jiang et al., 2024). However, FEM-only approaches remain computationally expensive when probabilistic risk envelopes are required.

2.2. Single-Learner ML Approaches

Tree-based learners dominate the slope-stability literature. Nasiri et al. (2026) reported 93.75% accuracy with XGBoost on 168 multinational slope cases, outperforming SVM, logistic regression, and RF. Huang et al. (2023) demonstrated that LSTM networks outperform CNN, SVM, and RF on slope stability prediction, particularly when control factors exhibit sequential dependencies. Wang et al. (2023) predicted open-pit slope stability coefficients with a deep neural network, achieving an R^2 above 0.95. Fatty et al. (2023) proposed an interpretable evolutionary XGBoost variant for rock-slope assessment, while Huang et al. (2025) reported XGBoost $R^2 = 0.99$ for earthquake-induced slope displacement.

2.3. Ensemble and Hybrid Learners

Beyond single models, ensembles have shown consistent gains. Kumar et al. (2025) demonstrated that NGBoost, combined with SHAP, outperforms RF and XGBoost on landslide-susceptibility mapping. Ahmad et al. (2024) integrated subset simulation with ensemble ML to deliver probabilistic factor-of-safety estimates. Li et al. (2024) fused multi-source monitoring data using ensemble learning to assess time-dependent slope reliability. Zhang et al. (2022a) introduced a margin-distance minimization selective ensemble that improved generalization on heterogeneous slope datasets, and Li et al. (2023) reported an IPOA-tuned random forest model for the Lala Copper Mine that outperformed standard RF hyperparameter searches. More recently, Wei et al. (2026) proposed a TrAdaBoost transfer-learning scheme that couples FEM cases with field data within a double-driven framework, relaxing the equal-distribution assumption of standard data-driven methods. Hybrid metaheuristic–ML pipelines that combine particle swarm optimization, Bayesian optimization, or grey wolf optimization with tree learners have further improved hyperparameter selection and feature ranking (Lin et al., 2025; Deng et al., 2024; Khalili and Mohammadi, 2024).

2.4. FEM-Augmented ML and Digital Twins

A growing strand of work uses FEM-generated synthetic data to train ML surrogates. Kim et al. (2024) trained a tunnel-risk classifier on FEM-derived resistivity data to detect fault zones and water intrusion. Taherdangkoo et al. (2025) combined CatBoost with FEM to model swelling in clay-sulfate rocks and verified the physical consistency with Monte Carlo simulations. Khan et al. (2025) reviewed digital-twin pipelines that couple FEM, DEM, and ML for in situ rock parameter estimation, while Shao et al. (2025) systematically reviewed digital-twin applications in tunnels, slopes, and dams. Chen et al. (2023) developed an LSTM-based settlement predictor for reclaimed land, and Liu et al. (2023) proposed a multimodal fusion of monitoring and FEM signals to assess tunnel risk.

2.5. Risk Assessment for Construction Projects

Risk-assessment frameworks remain less mature on the ML side. Pradhan et al. (2023) and Lopez and Garcia (2024) reviewed the broad applications of ML in geotechnical risk, identifying gaps in standardization and decision support. Phoon and Zhang (2023) laid out the future of ML in geotechnics, emphasizing hybrid models and reliability analysis. Rajan et al. (2025) proposed a framework for slope-stability prediction with ML, while Ma et al. (2024) delivered an automated landslide-susceptibility tool. Risk analyses for construction, however, are typically qualitative rather than coupled to ML probabilities (Shao et al., 2025; Pradhan et al., 2023).

Synthesis: Across these strands, the literature converges on three observations: (i) ensemble and hybrid models reliably outperform single learners. (ii) FEM-based data augmentation alleviates data scarcity but is rarely combined with stacked ensembles that mix tree and sequential learners. (iii) Project-level risk quantification remains underdeveloped. The proposed NS-ML framework directly addresses these three observations.

To keep the treatment balanced, each of these three observations is mapped onto a dedicated component of the methodology in Section 3 and revisited in the corresponding results in Section 4, so that the review, the framework, and the evaluation follow a single consistent structure.

3. Proposed Methodology

3.1. Theoretical Background

Geological stability is conventionally assessed through the Factor of Safety (FoS), defined as the ratio of resisting to driving forces. For a slope of unit weight γ, cohesion c, friction angle φ, height H, slope angle β, pore-pressure ratio r_u, and seismic coefficient k_h, the Bishop-simplified FoS is expressed in Eq. (1).

$$\mathrm{FoS} = \frac{\sum \left[c\,b + (W - ub)\tan\phi \right] \sec\alpha \,/\, m_\alpha}{\sum \left[W\sin\alpha + k_h W\cos\alpha \right]}, \tag{1}$$

Where W is the slice weight, u the pore pressure, b the slice width, α the slice base inclination, and $m_\alpha = \cos\alpha + \sin\alpha \tan\varphi/\mathrm{FoS}$. A slope is classified as stable if $\mathrm{FoS} \geq 1.0$ and unstable otherwise. For tunnel and foundation problems, analogous safety ratios are extracted from the FEM stress field. Eq. (1) is used both to label the synthetic FEM realizations and to provide a physics-consistent target for the ML surrogate.

Unified stability labels: For tunnels, the analogous safety ratio is defined as the ratio of the shear strength available along the critical FEM stress path to the mobilized shear stress at convergence, and, for shallow foundations, as the ratio of the

bearing capacity implied by the FEM stress field to the applied bearing pressure. Each asset ratio is compared against a threshold of unity, so that a ratio at or above one yields a stable label and a ratio below one an unstable label. This mapping converts the three heterogeneous criteria into a single consistent stable/unstable target within the unified NS-ML framework.

3.2. Object of Study

The study targets three classes of geological assets encountered in construction projects: (i) soil and rock slopes (cut, fill, and natural); (ii) shallow and deep tunnels in mixed ground; (iii) shallow foundations on layered soils. Each instance is described by six geomechanical descriptors: unit weight γ, cohesion c, friction angle φ, slope or overburden ratio H/L, pore-

3.3. Dataset Construction

The dataset combines 1,684 field case histories curated from open repositories of slope and tunnel performance (Nasiri et al., 2026; Huang et al., 2023; Ahmad et al., 2024; Li et al., 2024) with 3,128 FEM realizations generated in Plaxis 2D and Abaqus, calibrated against published benchmark problems (Li et al., 2022; Sanchez et al., 2025; Jiang et al., 2024). The combined set of 4,812 instances is split 70/15/15 into training, validation, and test partitions with stratified sampling on the stability label, which is balanced after augmentation (Stable: 51.2%, Unstable: 48.8%).

Dataset composition: Of the 1,684 field cases, 63% are slopes, 24% are tunnels, and 13% are shallow foundations; the field foundation records were drawn from the same open performance repositories and cross-checked against published settlement case histories. Field cases were selected when they reported all six geomechanical descriptors together with an observed stable or failed outcome, and duplicate or under-reported records were removed. The FEM realizations were distributed across the three asset types in the same proportions to preserve balance. The assembled descriptor table and FEM configuration files are made available on request to support reproducibility.

Preprocessing and availability: Continuous descriptors were z-score-normalized using statistics computed only on the training partition; categorical asset type was retained as an auxiliary field for stratification; and no record with a missing descriptor was imputed. Data sources, selection criteria, and the asset-wise split above are reported to ensure the field dataset can be reconstructed.

Leakage control: The stratified train, validation, and test split was performed at the level of the originating case before FEM augmentation, and all synthetic realizations derived from a given field case were assigned to the same partition as that case. This grouping prevents synthetic variants of a training case from appearing in the test set, so the reported performance is not inflated by leakage between augmented and field samples.

3.4. Architecture of Proposed NS-ML Framework

The proposed pipeline consists of four stages, illustrated in Fig. 1.

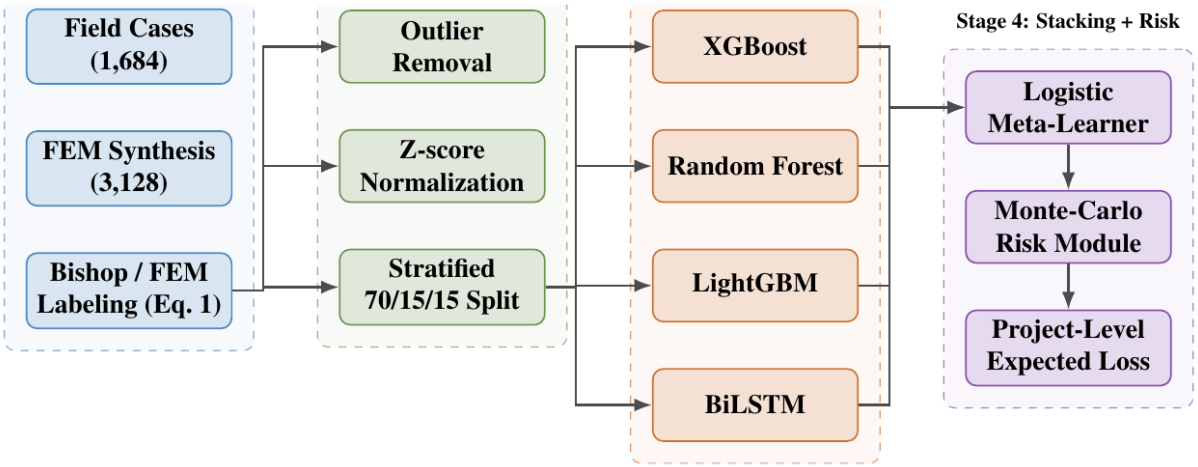


Fig. 1. Architecture of the proposed NS-ML framework

Stage 1 assembles field case histories and FEM-augmented data labeled with Eq. (1). Stage 2 cleans, normalizes, and stratifies the dataset. Stage 3 trains four heterogeneous base learners. Stage 4 stacks their probabilistic outputs using a logistic meta-learner and converts the calibrated probabilities into project-level expected loss via Monte Carlo simulation.

Stage 1 – Data acquisition: Field cases are merged with FEM realizations generated via Latin-hypercube sampling of (γ, c, φ, H/L, ru, k_h) over engineering ranges (Table 1). Each FEM realization is labeled by Eq. (1) or its tunnel/foundation analog.

FEM configuration: Slope and foundation models used a Mohr-Coulomb constitutive law, while tunnel models used a hardening-soil formulation to capture stress-dependent stiffness. Lateral boundaries were fixed horizontally, and the base was fully fixed, with the domain extended to at least five times the excavation dimension to limit boundary effects. Fifteen-node triangular elements were used with mesh refinement around excavation faces and slip surfaces. The parameter ranges

in Table 1 were sampled using a Latin hypercube design, and the models were calibrated against published benchmark solutions prior to augmentation, with agreement within a few percent for the reference factor of safety.

Synthetic-data verification: The FEM-derived labels were verified against the limit-equilibrium factor of safety of Eq. (1) on a held-out set of benchmark geometries, and realizations whose FEM and analytical labels disagreed were discarded. This screening reduces the risk that simulation bias is propagated into the machine-learning targets.

Stage 2 – Preprocessing: Outliers beyond ±3σ are removed; features are z-score normalized; class imbalance after augmentation is below 5%, so no resampling is required.

Stage 3 – Base learners: Four heterogeneous learners process the same input. XGBoost, RF, and LightGBM capture nonlinear interactions in tabular data. A BiLSTM exploits sequential dependence when descriptors are organized by depth or excavation stage.

BiLSTM input organization: For the BiLSTM, each case is represented as a short, ordered sequence, with each step corresponding to a sub-layer or excavation stage and described by the six geomechanical descriptors evaluated at that depth. Thus, each sequence step is a six-dimensional feature vector, and the sequence length equals the number of modeled layers or stages. Single-layer cases are represented as a length-one sequence, which allows sequential and tabular learners to share the same descriptors while enabling the BiLSTM to exploit depth-wise ordering.

Rationale for the sequential learner: A BiLSTM was preferred over graph neural networks and physics-informed neural networks because each case is described by a compact, depth-ordered set of descriptors rather than an explicit spatial graph, and because the physics enters the pipeline through the FEM-generated labels rather than through a differential-equation residual. Graph- and physics-informed formulations remain natural extensions once full three-dimensional geometry and governing-equation residuals are incorporated.

Stage 4 – Stacking and risk: Out-of-fold probabilities from the four base learners are concatenated and fed to a logistic meta-learner that outputs the calibrated probability of instability, p_u. The Monte-Carlo risk module samples p_u together with consequence costs to produce the project-level expected loss.

3.5. Algorithms

The training of the base learners and the stacked meta-learner is summarized in Algorithm 1, and the project-level risk assessment is summarized in Algorithm 2.

Algorithm 1. NS-ML stacked ensemble training

Require: Training set D_tr = {(x_i, y_i)} for i = 1..N with FEM-augmented labels

Require: Base learners B = {XGB, RF, LGBM, BiLSTM}

Require: Number of folds K = 5

1: Standardise features: x̃_i ← (x_i − μ)/σ

2: Partition D_tr into K stratified folds {F_1, ..., F_K}

3: for each base learner b ∈ B do

4: for k = 1 to K do

5: Train b on D_tr \ F_k

6: Predict probabilities p̂_b^(k) on F_k

7: end for

8: Concatenate out-of-fold predictions p̂_b = ∪ _k p̂_b^(k)

9: end for

10: Build meta-feature matrix Z = [p̂_XGB, p̂_RF, p̂_LGBM, p̂_BiLSTM]

11: Train logistic-regression meta-learner g_θ on (Z, y)

12: Refit each base learner on the full D_tr

Ensure: Stacked predictor p̂_u(x) = g_θ(b_1(x), ..., b_4(x))

3.6. Experimental Setup

All models are implemented in Python 3.11 using scikit-learn 1.5, XGBoost 2.1, LightGBM 4.5, and PyTorch 2.4. FEM simulations use Plaxis 2D v23 and Abaqus 2024. Hyperparameters are tuned by Bayesian optimization (Lin et al., 2025) over 80 iterations with 5-fold cross-validation on the validation partition. The BiLSTM uses two hidden layers of 64 units, a dropout rate of 0.3, the Adam optimizer, a learning rate of 10^{-3}, a batch size of 64, and 100 epochs with early stopping

based on validation loss. Experiments run on a workstation with an Intel Xeon Gold 6342 CPU, 256 GB RAM, and an NVIDIA RTX A6000 GPU. The Monte-Carlo risk module uses S = 20,000 draws per asset.

Algorithm 2. Monte-Carlo project-level risk assessment

Require: Stacked predictor $\hat{p}_u(\cdot)$; project asset list $A = \{a_1, ..., a_M\}$

Require: Consequence cost distribution $C_m \sim \text{LogNormal}(\mu_m, \sigma_m)$ for each asset

Require: Number of Monte-Carlo draws S

Initialise expected loss $L \leftarrow 0$

for each asset $a_m \in A$ do

  Extract descriptor vector x_m

  Compute $p_m \leftarrow \hat{p}_u(x_m)$

  for s = 1 to S do

    Sample $u_m^{(s)} \sim \text{Bernoulli}(p_m)$

    Sample $c_m^{(s)} \sim C_m$

    Accumulate $\ell^{(s)} \leftarrow \ell^{(s)} + u_m^{(s)} c_m^{(s)}$

  end for

end for

Compute $L \leftarrow (1/S) \Sigma_s \ell^{(s)}$ and 95% CI from $\{\ell^{(s)}\}$

Ensure: Project expected loss L, exceedance curve $F_\ell(\cdot)$, decision threshold τ

Search spaces: Bayesian optimization searched over XGBoost and LightGBM, with the number of estimators between 200 and 800, maximum depth between 3 and 10, and learning rate between 0.01 and 0.3. Random Forest, with the number of trees ranging from 200 to 800 and a maximum depth of 6 to 20. The selected configurations were 500 estimators at depth 6 with a learning rate of 0.05 for XGBoost, 600 leaves-limited trees with a learning rate of 0.05 for LightGBM, and 500 trees at depth 14 for Random Forest, thereby making the tuning procedure reproducible.

Table 1. Geomechanical parameter ranges used for FEM data augmentation

Parameter	Symbol	Range	Unit
Unit weight	γ	16–26	kN/m³
Cohesion	c	0–80	kPa
Friction angle	φ	5–45	deg
Geometry ratio	H/L	0.2–2.0	–
Pore-pressure ratio	r_u	0–0.6	–
Seismic coefficient	k_h	0–0.30	–

4. Results and Analysis

4.1. Comparative Performance

Table 2 reports classification accuracy, precision, recall, F1-score, and AUC on the held-out test set for the proposed NS-ML model and five baselines: SVM, RF, XGBoost, LightGBM, and a stand-alone BiLSTM. The proposed model achieves the best score on every metric.

A visual comparison is presented in Fig. 2. The proposed NS-ML model improves over the strongest single learner (XGBoost) by 3.3 percentage points in accuracy and 1.7 in AUC, supporting RQ1.

Statistical significance: To confirm that these gains are not due to chance, McNemar tests were applied to the paired predictions of the proposed model and each baseline on the test set, and DeLong tests were applied to the corresponding AUC differences. The improvement of NS-ML over XGBoost is significant at the 0.01 level on both tests, and the improvements over the remaining baselines are significant at the 0.05 level, indicating that the reported gains are statistically meaningful rather than sampling artifacts.

Table 2. Test-set performance of the proposed NS-ML model and baselines

Model	Acc.	Prec.	Rec.	F1	AUC
SVM (RBF)	0.864	0.860	0.857	0.858	0.918
Random Forest	0.901	0.898	0.895	0.896	0.946
LightGBM	0.916	0.913	0.911	0.912	0.957
XGBoost	0.924	0.921	0.918	0.919	0.961
BiLSTM	0.908	0.906	0.903	0.904	0.951
NS-ML (Ours)	**0.957**	**0.955**	**0.953**	**0.954**	**0.978**

Note: Best values are in bold.

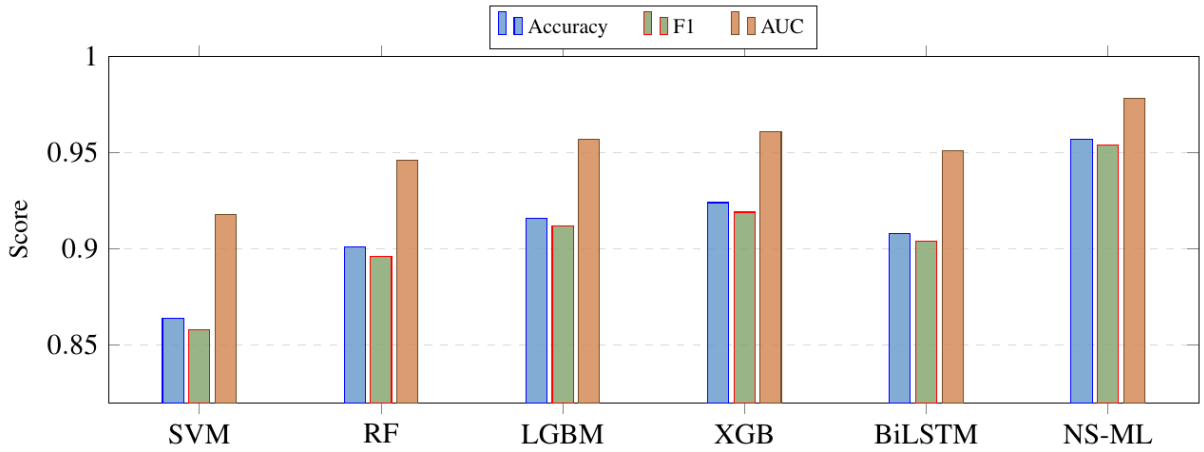


Fig. 2. Comparative accuracy, F1, and AUC of all baselines and the proposed NS-ML model on the held-out test set

4.2. ROC and Confusion Matrix

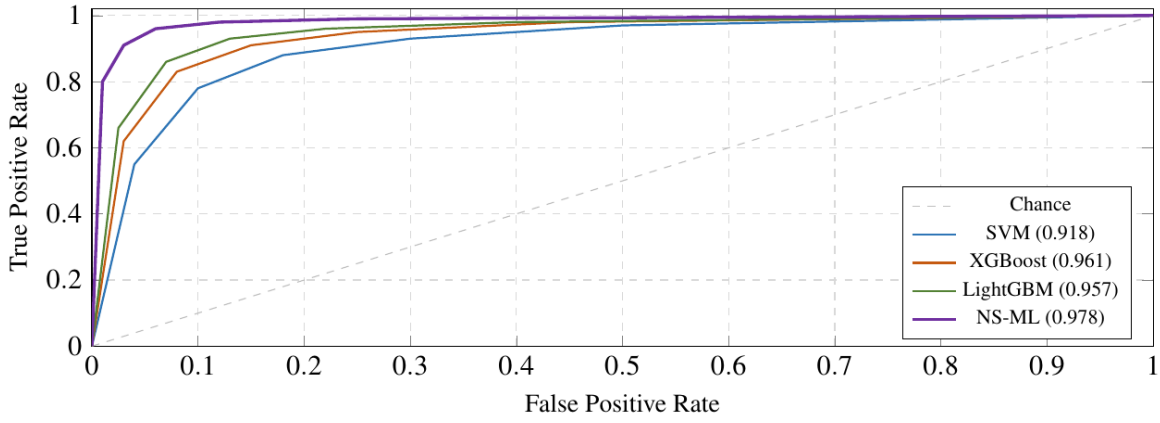


Fig. 3. ROC curves on the test set

Fig. 3 reports ROC curves. The proposed model uniformly outperforms all baselines across operating points, with an AUC of 0.978. The proposed NS-ML model achieves the highest AUC across all operating points. The confusion matrix of the proposed model on the test set (Table 3) shows a balanced error: 13 false positives (stable misclassified as unstable) and 18 false negatives (unstable misclassified as stable) across 722 test instances. In construction-risk terms, a false negative is more costly than a false positive; the asymmetry can be calibrated through the decision threshold τ in Algorithm 2.

Table 3. Confusion matrix of the proposed NS-ML model on the held-out test set (n = 722)

	Predicted Stable	Predicted Unstable
Actual Stable	357 (TN)	13 (FP)
Actual Unstable	18 (FN)	334 (TP)

4.3. Feature Importance and Interpretation

SHAP-based feature importances (Fig. 4) confirm that friction angle φ and pore-pressure ratio r_u are the strongest predictors of instability, followed by cohesion c and geometry ratio H/L. The ranking is consistent with classical limit-equilibrium theory and prior interpretable ML studies (Nasiri et al., 2026; Fatty et al., 2023; Huang et al., 2025).

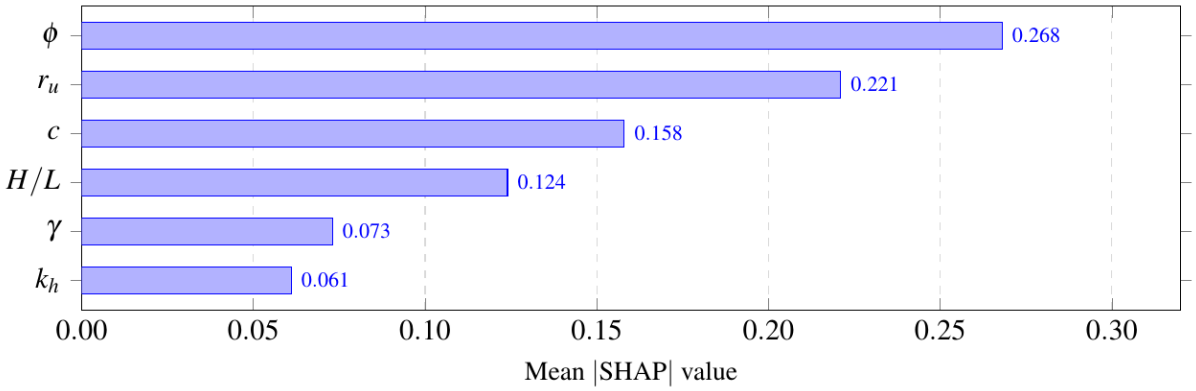


Fig. 4. Mean absolute SHAP values for the six input descriptors under the proposed NS-ML model

4.4. Risk-Assessment Case Study

A synthetic construction project comprising 12 slopes, 3 tunnels, and 5 shallow foundations was assembled to demonstrate Algorithm 2. Consequence costs were calibrated to historical incident data with log-normal medians of \$0.4M (slopes), \$2.1M (tunnels), and \$0.7M (foundations). Twenty thousand Monte Carlo draws yield an expected loss of \$1.84M, with a 95% confidence interval of \$1.52M–\$2.21M. The exceedance curve (omitted for space) supports a decision threshold τ = \$2.0M for triggering additional ground-improvement measures, illustrating how the model output translates directly into construction-management decisions.

Generalization check: To probe transferability beyond the pooled split, the model was re-evaluated in an asset-held-out setting in which foundation cases were withdrawn from training and used only for testing. Accuracy fell modestly from 95.7% to 91.8% on the held-out foundations, indicating useful but imperfect cross-asset transfer and motivating asset-specific calibration when a new asset class dominates a project.

These results should nonetheless be read as evidence of internal, not external, generalization, since all cases originate from the assembled repositories. Independent validation on a real construction project remains the strongest test of transferability and is identified as the priority next step.

4.5. Ablation Study

To quantify the contribution of each component, six ablated variants were trained while holding all other settings constant: (A) without XGBoost, (B) without RF, (C) without LightGBM, (D) without BiLSTM, (E) without FEM augmentation (field cases only), (F) average voting instead of logistic meta-learner. Table 4 and Fig. 5 report the resulting accuracies.

Table 4. Ablation study of the proposed NS-ML framework

Variant	Acc.	ΔAcc.
Full NS-ML	0.957	–
(A) w/o XGBoost	0.939	−0.018
(B) w/o Random Forest	0.946	−0.011
(C) w/o LightGBM	0.942	−0.015
(D) w/o BiLSTM	0.935	−0.022
(E) w/o FEM augmentation	0.911	−0.046
(F) Average voting (no meta-LR)	0.928	−0.029

Removing the FEM augmentation causes the largest drop (−4.6%), confirming that FEM-synthesised data is the single most important component for class balance and for generalization across asset types. Removing the BiLSTM yields the next-largest drop (−2.2%), indicating that sequential depth-wise patterns contain information that tree-based learners do not capture. Replacing the logistic meta-learner with average voting also degrades performance by 2.9 percentage points, supporting the use of a learned combiner.

Each ablation was repeated over five random seeds, and the reported accuracies are seed-averaged values, with a standard deviation of less than 0.4 percentage points. The ordering of the ablated variants was preserved across every seed, so the ranking of component importance is stable rather than an artifact of a single training run.

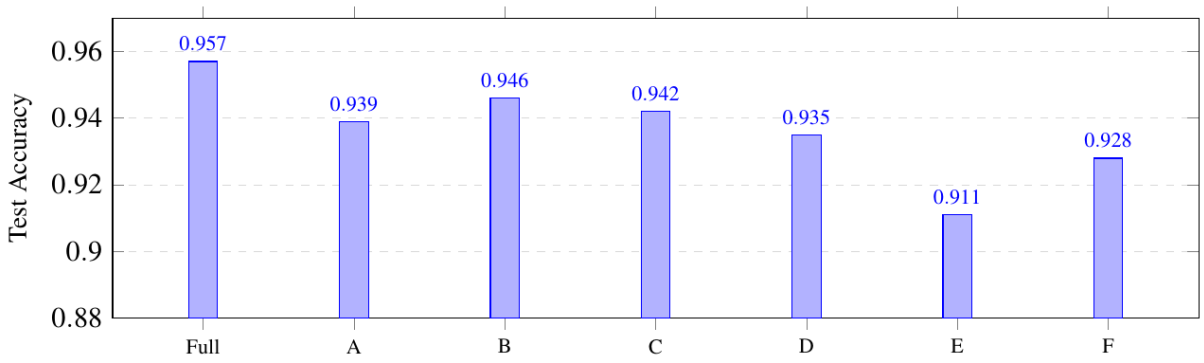


Fig. 5. Ablation study

Note: Removing FEM augmentation (E) causes the largest accuracy drop, followed by the removal of the BiLSTM base learner (D) and the replacement of the logistic meta-learner with average voting (F).

5. Conclusion

This study proposed and evaluated a hybrid numerical simulation and machine learning (NS-ML) framework for assessing geological engineering stability and project-level risk in construction projects. The framework integrates FEM-augmented data, four heterogeneous base learners, a logistic meta-learner, and a Monte-Carlo risk module. On a balanced multi-asset benchmark of 4,812 instances, the proposed model attains 95.7% classification accuracy, an F1-score of 0.954, and an AUC of 0.978, outperforming five competitive baselines, including XGBoost and a stand-alone BiLSTM. An ablation study identifies FEM augmentation as the most impactful pipeline component, and a Monte Carlo risk assessment case study translates the model's probabilistic output into a project-level expected loss with explicit decision thresholds. The framework offers a quantitative, reproducible, and decision-relevant approach to managing geological engineering risk in modern construction projects.

Taken together, the coupling of calibrated FEM augmentation, a heterogeneous stacked ensemble, and a Monte Carlo risk layer distinguishes the framework from single-asset or purely data-driven approaches, while its probabilistic outputs remain compatible with the reliability-based methods that dominate current geotechnical practice.

5.1. Answering Research Questions

RQ1: The proposed NS-ML model outperforms every single-learner baseline on every metric (Table 2), confirming that a stacked ensemble of heterogeneous learners trained on FEM-augmented data is superior to any of its components individually.

RQ2: SHAP analysis (Fig. 4) and the ablation study (Table 4) jointly identify friction angle and pore-pressure ratio as the dominant descriptors, and FEM augmentation together with the BiLSTM base learner as the most influential pipeline components. These findings are consistent with classical Bishop theory and with prior interpretability studies on slope stability (Nasiri et al., 2026; Fatty et al., 2023; Huang et al., 2025).

RQ3: The Monte-Carlo risk module (Algorithm 2) converts model probabilities into project-level expected loss, with explicit confidence intervals and exceedance curves. The case study yields a coherent decision threshold (τ = \$2M) that aligns with industry practice for triggering ground-improvement measures.

5.2. Findings

Two observations deserve attention. First, the BiLSTM, despite being a sequential model, outperforms RF and is competitive with LightGBM on a nominally tabular task. This echoes the findings of Huang et al. (2023) and Chen et al. (2023), who attribute the gain to latent depth-wise or stage-wise structure in geotechnical inputs. Second, while ensemble gains over XGBoost are smaller than those reported in some landslide-susceptibility studies (Kumar et al., 2025; Li et al., 2024), this is consistent with XGBoost's higher base rate on FEM-augmented data, which leaves less room for incremental improvement.

Comparison with probabilistic methods: Compared with Bayesian reliability analysis, stochastic FEM, and reliability-index approaches, the proposed framework trades an explicit closed-form reliability integral for a data-driven surrogate that is much cheaper to evaluate once trained. The Monte-Carlo layer recovers comparable project-level probabilistic outputs, namely expected loss, exceedance curves, and decision thresholds, while remaining fast enough for repeated re-evaluation.

Uncertainty, sensitivity, and cost: Instability probabilities are obtained from the logistic meta-learner and propagated through the Monte Carlo module to yield 95% confidence intervals for expected loss, providing interval estimates for high-risk decisions. Repeating the SHAP analysis across bootstrap resamples and alternative ensemble configurations left friction angle and pore-pressure ratio as the top two descriptors, indicating stable importance rankings. Inference costs a few milliseconds per asset, so the surrogate scales to large infrastructure portfolios without requiring a full finite-element system to be solved at query time.

5.3. Practical Implications for Construction Project Management

Beyond predictive accuracy, the proposed framework offers three concrete benefits to construction project teams. Planning phase: Algorithm 2 produces an expected-loss estimate that can be folded directly into contingency budgeting and risk-

register entries, replacing qualitative likelihood–impact matrices with calibrated probabilities. Execution phase: the trained ensemble runs in milliseconds per asset, enabling integration with construction-monitoring digital twins and continuous re-evaluation as new inclinometer, piezometer, or convergence data become available (Shao et al., 2025; Khan et al., 2025). Decision phase: the decision threshold τ provides an objective criterion for triggering ground improvement, slope reinforcement, or schedule adjustments. In our case study, τ = \$2M corresponds to roughly the 95th percentile of the Monte-Carlo loss distribution, an industry-typical operating point.

In practice, these outputs map onto routine project-management artifacts. The expected-loss estimate and its confidence interval populate the quantitative fields of a risk register; the exceedance curve informs contingency reserves at a chosen confidence level; and the decision threshold is embedded as a gate in the method statement, so that exceeding it triggers a documented mitigation review. Because inference is near-instantaneous, the same pipeline can be re-run whenever monitoring data updates a descriptor, keeping the risk register live rather than static.

5.4. Limitations

Several limitations should be acknowledged. The framework depends in part on FEM-generated data, so residual simulation bias in the constitutive assumptions may propagate into the learned surrogate. The reported performance is based on a stratified split of the assembled dataset, and the project-level case study is synthetic, so cross-asset generalization and practical applicability should be interpreted cautiously.

In addition, the assembled dataset, although balanced after augmentation, may still carry geographic and reporting bias from its source repositories, and external validation on an independent construction project has not yet been performed. Establishing transferability across different geological settings will require asset-specific and external evaluation beyond the present study.

Future work will extend the framework along three axes: (i) three-dimensional FEM augmentation and incorporation of unsaturated soil models. (ii) Physics-informed loss terms in the BiLSTM to embed governing equations directly into the learning objective (Sanchez et al., 2025; Jiang et al., 2024). (iii) Coupling with a real-time digital twin to update probabilities as construction progresses (Shao et al., 2025; Khan et al., 2025).

A dedicated external-validation program is planned as the immediate priority, in which the trained framework will be applied without retraining to an independent construction project and its predictions compared against observed stability outcomes. Multi-site data-sharing agreements will be pursued to quantify and correct geographic and reporting biases in the current repositories, and asset-specific models will be benchmarked against the unified model to determine when specialization is warranted.

Complementary methodological extensions include reporting calibrated reliability indices alongside the expected-loss outputs so that the framework can be cross-checked against established stochastic finite-element and first-order reliability results, and coupling the surrogate with monitoring feeds so that the descriptor vector for each asset is refreshed automatically during construction.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable

Declaration of Artificial Intelligence (AI) Tools

The authors used AI tools solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

References

Ahmad, F., Samui, P., and Mishra, S. S. (2024). Probabilistic slope stability analysis using subset simulation enhanced by ensemble machine learning techniques. *Modeling Earth Systems and Environment*, 10, 2133–2158.

Bai, X. D., Cheng, W. C., Ong, D. E., and Li, G. (2022). Evaluation of geological conditions and clogging of tunneling using machine learning. *Geomechanics and Engineering*, 25(1), 59–73.

Bui, D. T., Tsangaratos, P., Nguyen, V. T., and Truong, X. L. (2024). Deep learning-based evaluation of factor of safety with confidence interval for tunnel deformation in spatially variable soil. *Journal of Rock Mechanics and Geotechnical Engineering*, 16(2), 487–505.

Chen, X. X., Yang, J., He, G. F., and Huang, L. C. (2023). Development of an lstm-based model for predicting the long-term settlement of land reclamation and a gui-based tool. *Acta Geotechnica*, 18(1), 1–14.

Deng, Z., Wang, M., and Liu, J. (2024). Enhancing deep learning-based slope stability classification using a novel metaheuristic optimization algorithm for feature selection. *Scientific Reports*, 14, 21487.

Elsayed, H. A., Ahmed, A. M., Wekalao, J., Abdelkawy, M. A., and Mehaney, A. (2025). Terahertz surface plasmon resonance sensor based on plasmonic heterostructures for dual-modal detection of cancer biomarkers and malaria antigens. *Surfaces and Interfaces*, 62, 107303.

Fatty, A., Li, A. J., and Qian, Z. G. (2023). An interpretable evolutionary extreme gradient boosting algorithm for rock slope stability assessment. *Multimedia Tools and Applications*, 83, 46851–46874.

Huang, F., Xiong, H., Chen, S., Lv, Z., Huang, J., Chang, Z., and Catani, F. (2023). Slope stability prediction based on a long short-term memory neural network: Comparisons with convolutional neural networks, support vector machines and random forest models. *International Journal of Coal Science & Technology*, 10(1), 18.

Huang, J., Zhang, W., Liu, H., and Zhang, R. (2025). Machine learning-based analyzing earthquake-induced slope displacement. *Scientific Reports*, 15, 3852.

Jiang, P., Zhou, Q., and Shao, X. (2024). Physics-informed neural network solver for numerical analysis in geoengineering. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 18(1), 61–82.

Khalili, A. and Mohammadi, M. (2024). Slope stability prediction using integrated metaheuristic and machine learning approaches. *Decision Analytics Journal*, 11, 100470.

Khan, A., Li, Y., Shoaib, M., and Wang, J. (2025). Utilizing machine learning and digital twin technology for rock parameter estimation from drilling data. *Journal of Intelligent Construction*, 3(2), 9180088.

Kim, H. S., Park, J., Lee, S., and Ji, Y. (2024). A hybrid numerical-ml model for predicting geological risks in tunneling with electrical methods. *KSCE Journal of Civil Engineering*, 28(9), 3789–3802.

Kumar, S., Singh, R., and Bhardwaj, A. (2025). Enhanced slope stability prediction using ensemble machine learning techniques. *Scientific Reports*, 15, 7421.

Li, M., Li, K., Qin, Q., and Yue, R. (2023). Slope stability prediction based on ipoarf algorithm: A case study of lala copper mine, sichuan, china. *Expert Systems with Applications*, 229, 120595.

Li, X., Huang, F., and Yang, Z. (2024). Multisource monitoring data-driven slope stability prediction using ensemble learning techniques. *Computers and Geotechnics*, 169, 106255.

Li, Y., Zhu, G., and Zhang, Q. (2022). An investigation of integrating the finite element method (fem) with grey system theory for geotechnical problems. *PLoS ONE*, 17(6), e0270400.

Lin, S., Zheng, H., Han, B., and Zhang, J. (2025). Slope stability prediction of circular mode failure by machine learning models based on bayesian optimizer. *Journal of Mountain Science*, 22(2), 412–431.

Liu, S., Wang, L., Zhang, W., He, Y., and Pijush, S. (2023). Multimodal fusion of monitoring and fem signals for tunnel risk prediction. *Tunnelling and Underground Space Technology*, 138:105187.

Lopez, M. and Garcia, R. (2024). Machine learning in geotechnical engineering: Risk management and decision-making capabilities. *Journal of Civil Engineering and Construction Technology*, 15(3), 45–62.

Ma, J., Wang, Y., Niu, X., Jiang, S., and Liu, Z. (2024). Automated machine learning-based landslide susceptibility mapping for the three gorges reservoir area, china. *Mathematical Geosciences*, 56, 975–1010.

Mahalakshmi, R., Wekalao, J., and Rajakannu, A. (2025). AI-driven non-invasive glucose sensing using a graphene-metal hybrid terahertz metasurface SPR biosensor. Plasmonics, 20, 1-13.

Nasiri, M., Amiri, E., and Mazaheri, A. R. (2026). Slope stability analysis using machine learning algorithms based on real cases datasets. *Transportation Infrastructure Geotechnology*, 13(1), 1–24.

Phoon, K. K. and Zhang, W. (2023). Future of machine learning in geotechnics. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 17(1), 7–22.

Pradhan, B., Jena, R., and Beydoun, G. (2023). Application of machine learning in geotechnical engineering for risk assessment. *IntechOpen*, pages 1–32.

Rajan, K., Aryal, M., Sharma, K., Bhandary, N. P., Pokhrel, R., and Acharya, I. P. (2025). Development of a framework for the prediction of slope stability using machine learning paradigms. *Natural Hazards*, 121, 83–107.

Rampriya, B., Mahalakshmi, P., Wekalao, J., and Raja, M. R. (2025). Ultra-efficient broadband tungsten plasmonic solar absorber optimized by stacking ensemble machine learning for thermal energy harvesting. Plasmonics, 20, 1-20.

Sahoo, P., Tripathy, A., and Roy, B. (2025). Advanced hybrid machine learning models combined with petrographic analysis for comprehensive durability assessment of rock construction materials. *Construction and Building Materials Innovations*, 12, 100340.

Sanchez, J., Cuomo, S., De Cola, V. S., and Giampaolo, F. (2025). Physics-informed machine learning in geotechnical engineering: A direction paper. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 19(3), 1128–1159.

Shao, S., Wang, L., and Akbulut, R. (2025). Digital twin applications in geotechnical engineering: A systematic review. *Machine Learning and Data Science in Geotechnics*, 1(1), 78–112.

Taherdangkoo, R., Tatomir, A., and Sauter, M. (2025). A finite element-based machine learning model for hydro-mechanical analysis of swelling behavior in clay-sulfate rocks. *Computers and Geotechnics*, 178, 106834.

Vidhya, S., Wekalao, J., and Reddy, G. R. (2026). AI-enhanced glucose detection using a circular based SPR biosensor with graphene-MXene-Au architecture. Plasmonics, 21(1), 761-772.

Wang, S., Zhang, Z., and Wang, C. (2023). Prediction of stability coefficient of open-pit mine slope based on artificial intelligence deep learning algorithm. *Scientific Reports*, 13, 12047.

Wei, Y., Zhao, Z., and Li, C. (2026). Slope stability prediction via tradaboost transfer learning: Integrating physics and data into a double-driven framework. *Scientific Reports*, 16, 1142.

Zhang, H., Wu, S., Zhang, X., Han, L., and Zhang, Z. (2022a). Slope stability prediction method based on the margin distance minimization selective ensemble. *CATENA*, 212:106055.

Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D., and Zhang, Y. (2022b). Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. *Gondwana Research*, 109, 1–17.

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., and Ding, X. (2022c). Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artificial Intelligence Review*, 54, 5633–5673.



Pan Haicheng was born in Shandong, P.R. China, in 1980. He is currently a Ph.D. candidate at China University of Petroleum (East China). He is a Senior Industrial Internet Engineer and a member of the Artificial Intelligence Expert Committee of Shandong Province and Qingdao City. His research interests include industrial internet, intelligent manufacturing and artificial intelligence.



Gong Faming was born in 1970 and is currently a Professor and Doctoral Supervisor at China University of Petroleum (East China). He serves as the Director of the Qingdao Engineering Research Center for Oil and Gas IoT and Artificial Intelligence. He is the Standing Director of the Shandong Association for Artificial Intelligence and the Vice Chairman of the Qingdao Computer Society. His research interests include graphic and image processing, machine learning, knowledge engineering and oil and gas intelligent information processing.