Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(11), 4319-4338
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Gradient Boosting Surrogates for Limit-Equilibrium Factor of Safety of Mine Slopes with ICA Hyperparameter Tuning: A Duplicate-Free, Group-Aware Evaluation on a Compiled Dataset Relevant to Tarkwa, Ghana

Isaac Mensah Quansah[1,2*], Festus Kunkyin-Saadaari[1,2], Larry Testamen Karikari[1,2], Charles Asiedu[1,2]

[1]Department of Mining Engineering, Faculty of Mining and Minerals Technology, University of Mines and Technology (UMaT), Tarkwa, Ghana
[2]Festari Institute of Professional Studies, Tarkwa, Ghana
Email: *imensahquansah1482@gmail.com

How to cite this paper: Quansah, I.M., Kunkyin-Saadaari, F., Karikari, L.T. and Asiedu, C. (2026) Gradient Boosting Surrogates for Limit-Equilibrium Factor of Safety of Mine Slopes with ICA Hyperparameter Tuning: A Duplicate-Free, Group-Aware Evaluation on a Compiled Dataset Relevant to Tarkwa, Ghana. *Open Journal of Applied Sciences*, **16**, 4319-4338.
https://doi.org/10.4236/ojapps.2026.1611235

Received: September 21, 2026
Accepted: November 7, 2026
Published: November 10, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Accurate and rapid estimation of the Factor of Safety (FOS) of mine slopes supports geotechnical screening in open-pit mining. This study evaluates whether gradient boosting models can act as surrogates for Bishop's Simplified Method limit-equilibrium (LEM) FOS using seven input parameters (Rock Mass Rating, unit weight, cohesion, friction angle, slope angle, slope height and pore water pressure ratio). Three regularised learners (XGBoost, LightGBM, CatBoost) and three Imperialist Competitive Algorithm (ICA)-tuned counterparts were compared. The compiled dataset contained 281 records, of which only 103 were unique and 100 had unique input vectors; the remaining records were exact duplicates, which, when split at random, leak between training and test sets. After removing duplicates, records were grouped into 62 slope families (identical unit weight, cohesion, friction angle and slope height) and all validation kept families intact. Performance was assessed by repeated (2 × 5-fold) group-aware nested cross-validation and a final untouched test set of 13 families (20 records), with family-level bootstrap confidence intervals. Because FOS is itself computed from the model predictors, the models are LEM surrogates and were not validated against observed slope behaviour. Under this protocol, nested-CV R^2 was low: CatBoost R^2 = 0.27 (95% CI [0.11, 0.40],

RMSE = 0.34) and ICA-CatBoost R^2 = 0.23; only these two models improved significantly on a mean-only baseline (RMSE 0.39). No pairwise difference between CatBoost and ICA-CatBoost, or between any ICA model and its base learner, was statistically supported, and the ICA overfitting penalty did not improve held-out error relative to an unpenalised search. At the FOS = 1.0 threshold, nearly all unstable slopes were predicted as stable. Friction angle was the highest-ranked predictor in all six models, but with wide uncertainty. The results indicate that, for this dataset, the seven predictors do not support reliable FOS surrogate modelling, and that high R^2 values on the raw dataset are largely attributable to duplicate leakage.

Keywords

Slope Stability, Factor of Safety, Gradient Boosting, Imperialist Competitive Algorithm, Surrogate Modelling, Group-Aware Validation, Open-Pit Mine

1. Introduction

The stability of slopes in open-pit mining is one of the most critical geotechnical engineering challenges, with slope failures causing loss of life, equipment damage, production interruption and costly remediation [1] [2]. The Factor of Safety (FOS), defined as the ratio of available shear resistance to mobilised shear stress along a potential failure surface, is the universal quantitative index for slope stability assessment [3]. In West African gold mining, the Birimian-Tarkwaian geological sequence produces highly variable geotechnical conditions: weak, deeply weathered phyllites (cohesion 0 - 15 kPa, friction angle 15˚ - 25˚) coexist with competent Banket Series conglomerates (cohesion 25 - 80 kPa, friction angle 30˚ - 40˚) [4] [5]. This variability, combined with high annual rainfall (approximately 1800 mm) driving seasonal pore water pressure fluctuations, makes accurate and rapid FOS prediction technically challenging and operationally critical.

Conventional limit equilibrium methods (LEM), including Bishop's Simplified Method [6], the Morgenstern-Price Method [7] and Spencer's Method [8] (see also [9] [10]), remain the industry standard but require significant engineering resources for each analysis. Machine learning (ML) can provide fast surrogates of such calculations. The gradient boosting family, comprising XGBoost [11], LightGBM [12] and CatBoost [13], is widely used for tabular data. Two methodological weaknesses can, however, make reported accuracy on small geotechnical datasets unreliable: i) overfitting under default hyperparameters, and ii) information leakage occurs when duplicated or closely related records are split at random between training and test sets. Metaheuristic optimisation, including Particle Swarm Optimisation (PSO), Genetic Algorithms (GA) and the Imperialist Competitive Algorithm (ICA) [14], has been used to tune ML models in geotechnical applications [15]-[17]. Such searches typically minimise cross-validated error and can themselves overfit a small validation set.

This study makes the following contributions: 1) a duplicate-free, group-aware account of a compiled 281-record slope dataset, showing that only 103 records are unique and that random splitting leaks information; 2) a repeated nested cross-validation and final held-out evaluation of regularised XGBoost, LightGBM and CatBoost and ICA-tuned counterparts as surrogates for Bishop-method FOS, with confidence intervals and paired comparisons; 3) a diagnostic of the ICA overfitting penalty, including a comparison with an unpenalised search; and 4) decision-threshold (FOS = 1.0 and 1.5) performance and permutation importance. The compiled data are relevant to the Tarkwa mining region, Ghana; the models are LEM surrogates, not site-validated predictors.

2. Literature Review

2.1. Machine Learning for Slope Stability Prediction

ML-based slope stability prediction has progressed from early artificial neural network (ANN) applications in geotechnical engineering [18] and slope stability [19], including site-specific studies [20], to ensemble and metaheuristic-hybrid approaches [21]. Published performance figures are not directly comparable because datasets, outputs (FOS or stable/unstable class) and validation protocols differ, and many studies report a single random split. A remaining gap is the systematic reporting of group-aware validation, uncertainty and decision-threshold errors for FOS surrogate models.

2.2. Gradient Boosting Algorithms

XGBoost [11] implements gradient boosting with L1/L2 regularisation [22], tree pruning and column/row subsampling. LightGBM [12] introduces gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) with leaf-wise tree growth. CatBoost [13] employs ordered boosting and symmetric trees to reduce gradient estimation bias, with l2_leaf_reg providing regularisation. These algorithms have not, to our knowledge, been compared under identical regularisation and group-aware validation for slope FOS surrogate modelling.

2.3. Imperialist Competitive Algorithm

ICA [14], inspired by socio-political imperialism, operates through assimilation (colonies moving toward imperialists), revolution (random perturbation), empire competition and collapse. ICA has been applied to engineering optimisation problems [23] and to geotechnical prediction tasks including tunnel boring machine (TBM) penetration rate [15], tunnel squeezing [16] and soil strength prediction [17]. The present study uses ICA with a penalised objective for gradient boosting hyperparameter search and tests whether the penalty adds value over unpenalised search.

2.4. Geotechnical Context: Tarkwa, Ghana

The Tarkwa mining district hosts multiple large gold operations within the Bi-

rimian-Tarkwaian sequence. [4] reported cohesion of 5 - 35 kPa and friction angles of 15˚ - 32˚ for lateritic soils in the Tarkwa area. Pore water pressure is commonly an important time-variable control on slope stability in tropical open-pit mines, and is represented here by the pore pressure ratio. To our knowledge, no ML-based slope stability study specific to this geological context has been published.

3. Materials and Methods

3.1. Dataset

The dataset comprises 281 slope case records compiled from Iduapriem Mine investigations, published geotechnical literature and numerical simulations. Seven input parameters (RMR, unit weight, cohesion, friction angle, slope angle, slope height and pore water pressure ratio) and one output (FOS, computed by Bishop's Simplified Method [6]) are recorded per record. RMR is used as supplied in the compiled dataset (values 0.69 - 0.98, *i.e.* a normalised rating rather than the conventional 0 - 100 scale [24]). The unit of analysis is a unique slope-parameter record (a unique combination of the seven inputs). The compiled file does not carry a per-record source identifier; the numbers of records originating from Iduapriem investigations, published cases and numerical simulations are therefore not reported, and source-level validation was not possible (Section 4.8). Duplicate and related records were identified as follows (**Table 1**): records with identical values for all seven inputs and FOS were treated as exact duplicates and removed (178 of 281 records); the three input vectors that appeared with different FOS values (for example identical inputs with FOS = 0.80 and 1.86) were merged by averaging FOS; and the remaining 100 unique input vectors were grouped into 62 slope families, defined as records sharing identical unit weight, cohesion, friction angle and slope height (differing only in slope angle, pore pressure ratio and/or RMR). Families were never split between partitions, which also acts as a conservative proxy for records from the same source or simulation series. Identical inputs yielding different FOS values indicate that some inputs governing the LEM calculation (for example detailed geometry or slip-surface settings) are not captured by the seven recorded predictors.

FOS generation. FOS was computed for every record with Bishop's Simplified Method [6], a limit-equilibrium method for circular slip surfaces. The compiled dataset stores the resulting FOS and the seven inputs, but not the analysis settings (software, slip-surface search, number of slices, slope geometry beyond slope angle and height, or groundwater model beyond the pore pressure ratio ru, which is the only groundwater descriptor available). These settings could therefore not be reported or varied. Because FOS is a deterministic function of the same inputs used as model predictors, the machine-learning models are LEM surrogates, and no independent comparison with observed slope performance was made. Descriptive statistics of the 100 unique input vectors are summarised in **Table 2**.

Table 1. Source-by-source account and duplicate handling for the 281 compiled records.

Stage	Records	Note
Raw compiled records	281	Iduapriem investigations, published cases and numerical simulations; per-record source labels are not retained in the compiled file
Exact duplicate records removed	178	Identical in all seven inputs and FOS
Unique records	103	Three input vectors carried conflicting FOS values (six records)
Unique input vectors (analysis dataset)	100	Conflicting FOS averaged
Slope families	62	Same unit weight, cohesion, friction angle and slope height; kept intact in all splits
Development set/final test set	80/20	49/13 families; stratified by FOS class (<1.0, 1.0 - 1.5, ≥1.5), random seed 42

Table 2. Descriptive statistics of the analysis dataset (100 unique input vectors).

Parameter	Symbol	Min	Max	Mean	SD	Unit
Rock Mass Rating	RMR	0.690	0.980	0.850	0.090	-
Unit Weight	γ	12.00	28.44	20.18	3.12	kN/m^3
Cohesion	C	0.00	150.05	23.46	31.21	kPa
Friction Angle	φ	0.00	45.00	26.78	11.50	°
Slope Angle	β	16.00	53.00	32.89	9.81	°
Pore Pressure Ratio	Ru	3.66	214.00	44.54	48.75	-
Slope Height	H	0.000	0.500	0.249	0.179	m
Factor of Safety	FOS	0.625	2.310	1.298	0.400	-

SD = standard deviation.

In the analysis dataset, 19 records (19%) have FOS < 1.0 (unstable), 56 (56%) have 1.0 ≤ FOS < 1.5 (marginal) and 25 (25%) have FOS ≥ 1.5 (stable). Figure 1 presents the Pearson correlation heatmap.

3.2. Regularised Base Learner Configurations

All three base learners were configured with explicit regularisation (Table 3). The configurations adopted are as follows.

XGBoost: n_estimators = 300, learning_rate = 0.05, max_depth = 4, subsample = colsample_bytree = 0.80, reg_alpha = 0.50, reg_lambda = 2.00, min_child_weight = 5, gamma = 1.0.

LightGBM: n_estimators = 300, learning_rate = 0.05, max_depth = 5, num_leaves = 20, subsample = colsample_bytree = 0.80, reg_alpha = 0.50, reg_lambda = 2.00, min_child_samples = 15.

CatBoost: iterations = 300, learning_rate = 0.03, depth = 5, l2_leaf_reg = 8.0,

rsm = 0.80, Bernoulli bootstrap with subsample = 0.80 (no early stopping). Random seed = 0 for all models (xgboost 3.4.1, lightgbm 4.7.0, catboost 1.2.10, scikit-learn 1.8.0).

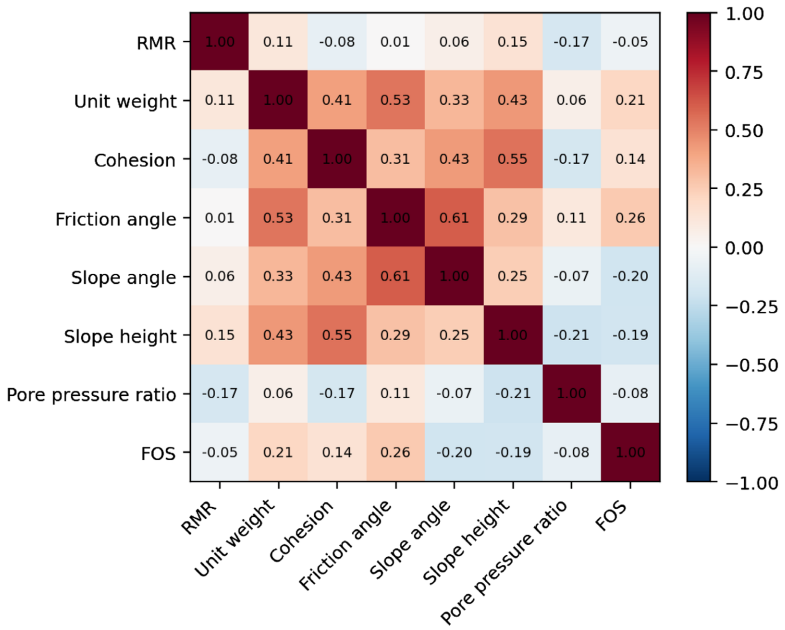


Figure 1. Pearson correlation heatmap of the analysis dataset. Friction angle (r = 0.26 with FOS), unit weight (r = 0.21), slope angle (r = −0.20) and slope height (r = −0.19) are the strongest linear correlates; no predictor correlates strongly with FOS.

3.3. ICA Hybridisation with Anti-Overfitting Objective

3.3.1. Anti-Overfitting Objective Function

The ICA objective augments cross-validated RMSE with an overfitting penalty, as given in Equation (1):

$$f(\theta) = \mathrm{RMSE}_{\mathrm{CV}} + \lambda \cdot \max\left(0, \mathrm{RMSE}_{\mathrm{CV}} - \mathrm{RMSE}_{\mathrm{train}}\right) \quad (1)$$

where $\mathrm{RMSE}_{\mathrm{CV}}$ is the 3-fold group-aware cross-validated RMSE on the training portion (families kept intact), $\mathrm{RMSE}_{\mathrm{train}}$ is the RMSE of the model refitted on that whole training portion, and $\lambda = 0.3$. The penalty is positive when generalisation error exceeds training error. A penalty of the form max (0, $\mathrm{RMSE}_{\mathrm{train}} - \mathrm{RMSE}_{\mathrm{CV}}$) would be identically zero in practice, because training error is almost always below cross-validated error. The coefficient $\lambda = 0.3$ was fixed a priori and was not tuned; its effect is examined by comparing with $\lambda = 0$ (Section 4.1).

3.3.2. ICA Algorithm and Configuration

ICA was configured with 12 countries, 3 imperialists, 20 iterations and a revolution rate of 0.3. Hyperparameter vectors were scaled to [0, 1] and decoded to the ranges in Table 3 (integers rounded). Assimilation moves each colony toward an assigned imperialist; revolution randomly repositions 30% of moves; the search retains the best-so-far solution. (Empire competition and collapse were not implemented in this simplified ICA.) The ICA seed was 10 × repeat + fold in nested

CV and 42 for the final fit. Table 3 presents the search spaces and the optimal hyperparameters identified in the final fit on the development set.

Table 3. Regularised base configurations and ICA-identified hyperparameters (final fit on the 80-record development set).

Model	Parameter	Search range	ICA optimum	Regularised base
ICA-XGBoost	learning_rate	[0.01, 0.15]	0.0462	0.05
	max_depth	[3, 6]	3	4
	n_estimators	[100, 400]	250	300
	reg_alpha	[0.1, 2.0]	0.1	0.5
	reg_lambda	[1.0, 5.0]	1	2.0
	min_child_weight	[3, 10]	4	5
ICA-LightGBM	learning_rate	[0.02, 0.10]	0.02	0.05
	num_leaves	[8, 25]	25	20
	min_child_samples	[10, 30]	30	15
	reg_alpha	[0.5, 3.0]	3	0.5
	reg_lambda	[1.5, 6.0]	4.32	2.0
ICA-CatBoost	learning_rate	[0.02, 0.10]	0.0213	0.03
	depth	[3, 6]	3	5
	l2_leaf_reg	[4.0, 12.0]	4.76	8.0

Optimal values are for the $\lambda = 0.3$ objective; values from nested-CV folds differ.

3.4. Validation Protocol and Evaluation Metrics

The 100 analysis records were first split into a final test set (20 records, 13 slope families) and a development set (80 records) using StratifiedGroupKFold (5 splits, shuffled, random_state = 42; first split), stratified by FOS class and grouped by slope family. The test set was used once, after model selection. On the development set, repeated nested cross-validation was performed: the outer loop used StratifiedGroupKFold (5 folds, random_state = 100 and 101 for two repeats, 10 outer folds); inside every outer training portion the ICA search used 3-fold GroupKFold. Metrics are pooled out-of-fold predictions per repeat, averaged over repeats; 95% confidence intervals come from 1,000 bootstrap resamples of slope families. Paired comparisons use the difference in RMSE with a family-bootstrap interval, a Wilcoxon signed-rank test on the 10 outer-fold RMSEs, and the Nadeau-Bengio corrected resampled t-test. P-values are uncorrected for multiplicity and are exploratory. Seven metrics were computed (R, R^2, RMSE, MSE, MAE, $VAF = \left[1 - \mathrm{Var}\left(y - \hat{y}\right)/\mathrm{Var}\left(y\right)\right] \times 100\%$, MAPE). A mean-only predictor

(training-fold mean FOS) provides a reference. Decision-threshold performance was evaluated at FOS = 1.0 and 1.5 (unstable predicted as stable = missed unstable).

3.5. Sensitivity Analysis

Permutation feature importance [25] was computed on each outer test fold (10 permutations per feature; mean decrease in R^2 averaged over the 10 outer folds) and on the final test set (20 permutations). Importance values are meaningful only where the model has predictive skill.

4. Results and Discussion

4.1. ICA Convergence

Figure 2 shows the best-so-far penalised objective for the final ICA runs. The objectives decline over the first ~10 iterations and are flat thereafter, with ICA-CatBoost and ICA-LightGBM reaching about 0.335 - 0.338; the curves show that the search converged on the training portion but do not show that held-out performance improved. The penalty term was active: in all 10 outer folds the penalty was non-zero at the selected solution for all three learners (mean 0.019, 0.043 and 0.188 RMSE units for XGBoost, LightGBM and CatBoost) and non-zero for 100% of candidate solutions. Table 4 compares the penalised search ($\lambda = 0.3$) with an unpenalised search ($\lambda = 0$) using the same seeds. The penalty reduced the gap between cross-validated and training RMSE of the selected solution but did not reduce held-out-fold error: the penalised search had equal or slightly higher outer-fold RMSE than the unpenalised search. The study therefore provides no evidence that the proposed objective contributes beyond ordinary cross-validated error minimisation. The choice of $\lambda = 0.3$ was not tuned or justified empirically.

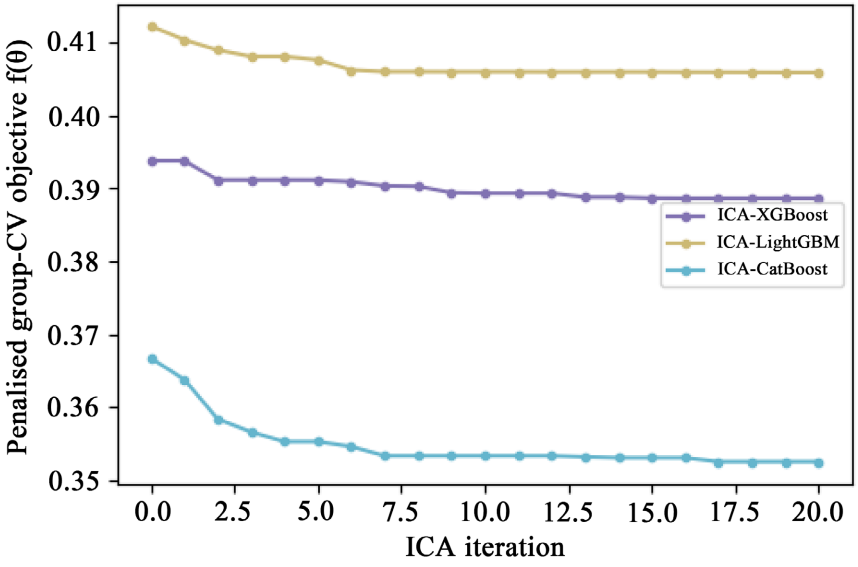


Figure 2. Best-so-far penalised group-CV objective for the final ICA runs (development set, $\lambda = 0.3$).

Table 4. Overfitting-penalty diagnostic and $\lambda = 0$ ablation (nested CV, 10 outer folds). Positive differences mean the penalised search had higher error.

Model	Penalty at selected solution (mean; non-zero folds)	Candidates with non-zero penalty	Outer-fold RMSE, $\lambda = 0.3$	Outer-fold RMSE, $\lambda = 0$	Difference ($\lambda_{0.3} - \lambda_0$) [95% CI]
ICA-XGBoost	0.019; 10/10	100%	0.389	0.386	0.004 [−0.012, 0.019]
ICA-LightGBM	0.043; 10/10	100%	0.383	0.351	0.032 [−0.018, 0.072]
ICA-CatBoost	0.188; 10/10	100%	0.342	0.329	0.013 [0.001, 0.026]

4.2. Performance Metrics and Overfitting Analysis

Table 5 gives the primary nested cross-validation results with 95% confidence intervals and a mean-only reference; Table 6 gives the single untouched test evaluation. Figure 3 compares training, cross-validated and held-out RMSE.

Table 5. Repeated group-aware nested cross-validation results (80 development records, 2 × 5 folds). CIs: family bootstrap. Train RMSE: mean over outer-fold training fits.

Model	R^2	R^2 95% CI	RMSE	RMSE 95% CI	MAE	MAPE (%)	Train RMSE
XGBoost	−0.000	[−0.12, 0.06]	0.393	[0.329, 0.451]	0.328	27.0	0.347
LightGBM	0.155	[−0.12, 0.36]	0.361	[0.298, 0.408]	0.297	25.0	0.216
CatBoost	0.271	[0.11, 0.40]	0.336	[0.278, 0.386]	0.273	22.3	0.107
ICA-XGBoost	0.017	[−0.10, 0.06]	0.390	[0.325, 0.444]	0.325	26.8	0.376
ICA-LightGBM	0.046	[−0.11, 0.12]	0.384	[0.325, 0.436]	0.322	26.7	0.345
ICA-CatBoost	0.230	[0.04, 0.36]	0.345	[0.292, 0.393]	0.283	23.2	0.162
Mean predictor	−0.003	[−0.10, −0.00]	0.394	[0.325, 0.450]	0.323	26.6	-

Table 6. Final untouched test evaluation (20 records; models fitted on the 80-record development set).

Model	Phase	R	R^2	RMSE	MSE	MAE	VAF (%)	MAPE (%)	Δ
XGBoost	Train	0.649	0.202	0.351	0.123	0.283	20.2	23.4	-
	Test	0.244	0.025	0.404	0.163	0.341	4.9	31.9	0.176
LightGBM	Train	0.898	0.788	0.181	0.033	0.141	78.8	11.2	-
	Test	0.838	0.626	0.250	0.063	0.215	62.6	19.4	0.162
CatBoost	Train	0.977	0.922	0.110	0.012	0.088	92.2	7.3	-
	Test	0.598	0.339	0.333	0.111	0.248	34.2	23.1	0.583
ICA-XGBoost	Train	0.752	0.315	0.326	0.106	0.263	31.5	21.8	-
	Test	0.379	0.092	0.390	0.152	0.326	10.6	30.4	0.223

Continued

ICA-LightGBM	Train	0.433	0.109	0.371	0.138	0.303	10.9	25.0	-
	Test	−0.152	−0.102	0.430	0.185	0.366	−6.0	34.6	0.211
ICA-CatBoost	Train	0.938	0.809	0.172	0.030	0.139	80.9	11.4	-
	Test	0.648	0.372	0.325	0.105	0.262	37.4	24.2	0.437

$\Delta = R^2_{train} - R^2_{test}$. With only 20 test records the 95% CIs are very wide (e.g. CatBoost R^2 0.34, CI [−0.39, 0.53]; LightGBM 0.63, CI [0.13, 0.71]); the held-out ranking differs from the nested-CV ranking and should not be used to select a model.

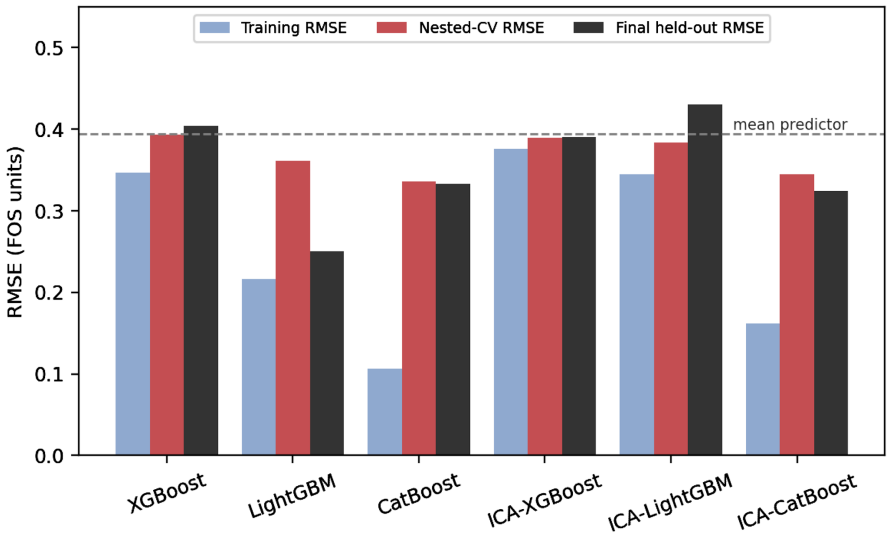


Figure 3. Training, nested-CV and final held-out RMSE for all six models. The dashed line is the nested-CV RMSE of the mean-only predictor.

Under nested group-aware CV, regularised CatBoost had the lowest RMSE (0.336, 95% CI [0.278, 0.386]) and highest R^2 (0.27, CI [0.11, 0.40]), followed by ICA-CatBoost (R^2 = 0.23). XGBoost, ICA-XGBoost and ICA-LightGBM were indistinguishable from the mean-only predictor (R^2 = −0.00), and LightGBM was not significantly better. Paired analysis against the mean-only predictor (**Table 7**) shows significant RMSE improvements only for CatBoost (ΔRMSE = 0.058, CI [0.023, 0.095]) and ICA-CatBoost (0.048, CI [0.013, 0.086]); absolute errors remain large (MAE ≈ 0.27 - 0.33 FOS units, MAPE ≈ 22% - 27%). The large gap between training and cross-validated RMSE for CatBoost (0.107 vs 0.336) indicates that the fitted models substantially overfit even with regularisation.

These results contrast sharply with the high R^2 values obtainable when duplicates are present: in a preliminary check on the original 281 records, random shuffled 5-fold CV of a generic gradient boosting model gave $R^2 \approx 0.87$, whereas duplicate-aware CV gave $R^2 \approx 0.62$, and 45 of the 57 test records of a random 80/20 split had an identical-input twin in the training set. Single-split R^2 values obtained from the raw dataset therefore overstate generalisation.

4.3. Predicted versus Actual FOS

Figure 4 and Figure 5 present pooled out-of-fold predicted versus observed FOS for the best standalone model (CatBoost) and the best ICA hybrid (ICA-CatBoost). Predictions are compressed toward the centre of the FOS range: low-FOS slopes are over-predicted and high-FOS slopes are under-predicted, consistent with limited skill.

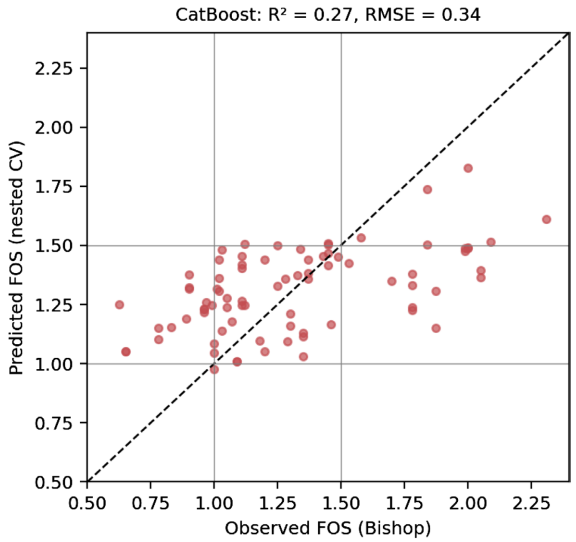


Figure 4. Predicted versus observed FOS for regularised CatBoost (pooled out-of-fold, nested CV).

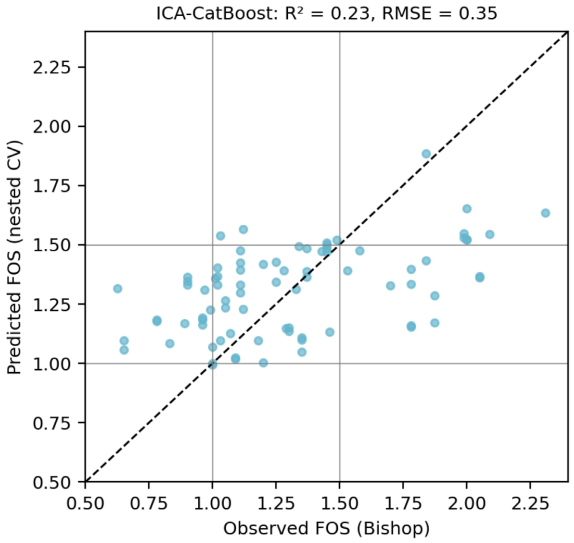


Figure 5. Predicted versus observed FOS for ICA-CatBoost (pooled out-of-fold, nested CV).

4.4. Error Distribution Analysis

Figure 6 shows residual distributions (pooled out-of-fold) for all six models and Figure 7 the cumulative distribution of absolute errors, including the mean-only

reference. CatBoost and ICA-CatBoost have somewhat narrower residual distributions than the other models; the improvement over the mean-only predictor is modest.

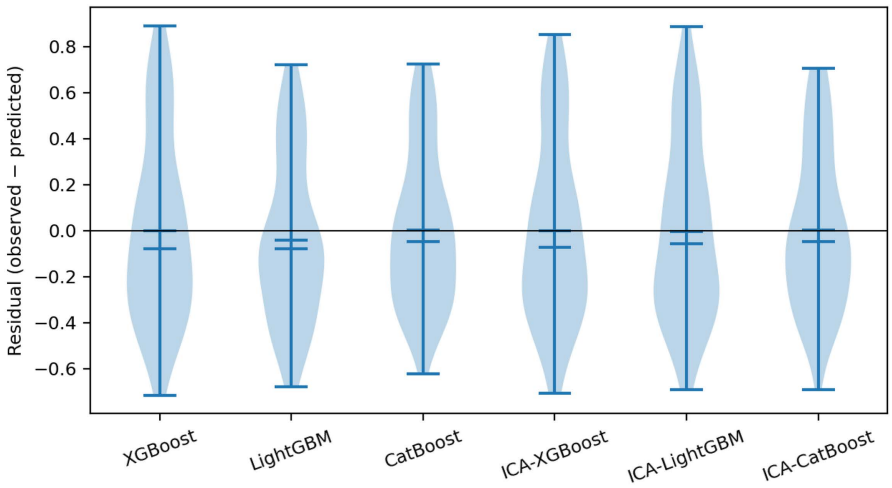


Figure 6. Violin plots of residuals (observed-predicted FOS), pooled out-of-fold.

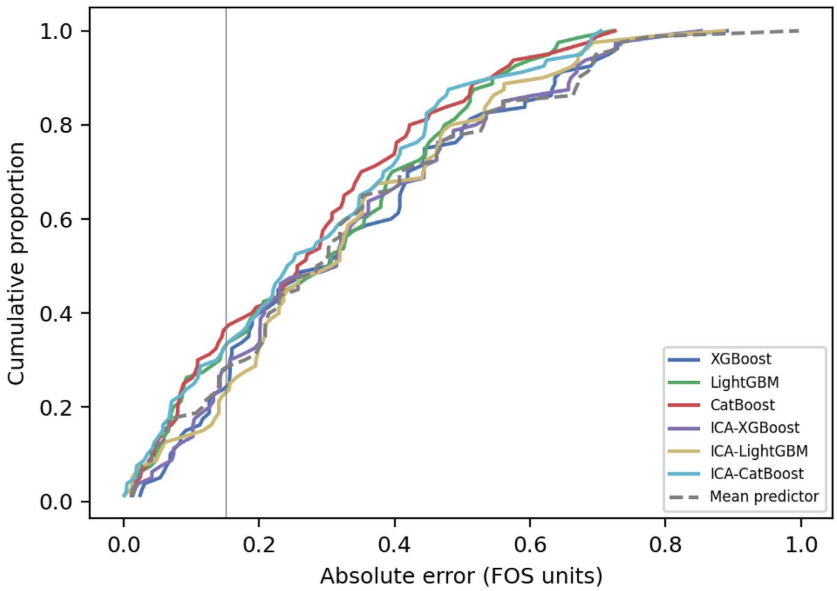


Figure 7. Cumulative distribution of absolute prediction errors (pooled out-of-fold).

4.5. Uncertainty, Statistical Comparison and Decision-Threshold Performance

Figure 8 shows RMSE with 95% family-bootstrap intervals for nested CV and the final test set; the intervals for most models overlap each other and the mean-only predictor. Table 7 reports paired comparisons. CatBoost has lower RMSE than XGBoost (ΔRMSE = −0.060, CI [−0.084, −0.032], corrected P = 0.019), but its advantage over LightGBM (−0.023, CI [−0.051, 0.002]) and over ICA-CatBoost (−0.009; corrected P = 0.49) is not supported. Figure 9 shows that ICA tuning did

not improve held-out RMSE for any base learner. "Best model" statements are therefore limited to CatBoost being the most consistent performer in nested CV; they are not supported by the held-out test, where LightGBM scored highest on 20 records.

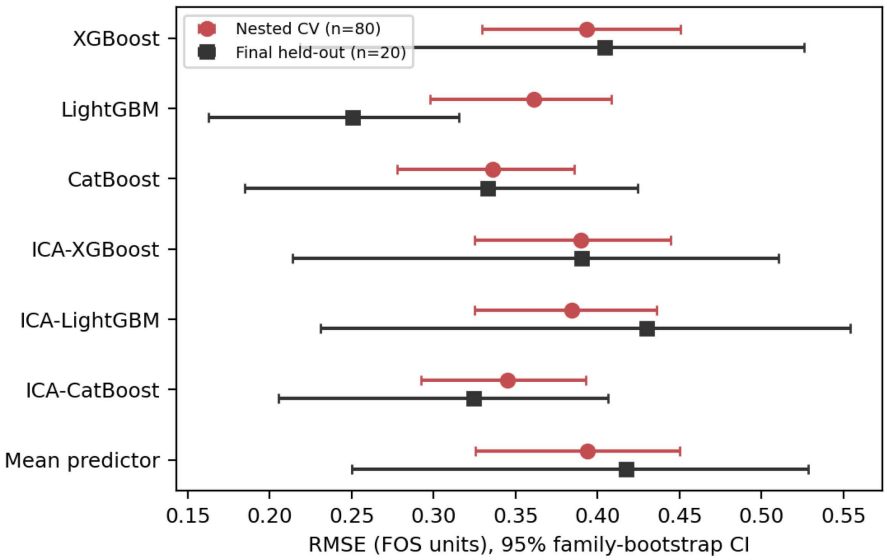


Figure 8. RMSE with 95% family-bootstrap CIs: nested CV (n = 80) and final held-out test (n = 20).

Table 7. Paired RMSE comparisons (nested CV, 10 outer folds). Negative ΔRMSE favours A. For "Mean predictor vs model", positive values favour the model. P-values are uncorrected and exploratory.

Comparison (A vs B)	ΔRMSE (A − B)	95% CI	Wilcoxon p	Corrected t-test p
CatBoost vs XGBoost	−0.060	[−0.084, −0.032]	0.002	0.019
CatBoost vs LightGBM	−0.023	[−0.051, 0.002]	0.037	0.204
CatBoost vs ICA-CatBoost	−0.009	[−0.019, −0.001]	0.193	0.491
LightGBM vs ICA-LightGBM	−0.026	[−0.069, 0.017]	0.375	0.502
XGBoost vs ICA-XGBoost	0.004	[−0.006, 0.013]	0.492	0.694
ICA-LightGBM vs ICA-XGBoost	−0.007	[−0.020, 0.008]	0.322	0.652
CatBoost vs ICA-LightGBM	−0.049	[−0.080, −0.017]	0.014	0.143
Mean predictor vs XGBoost	0.001	[−0.013, 0.014]	-	-
Mean predictor vs LightGBM	0.033	[−0.017, 0.085]	-	-
Mean predictor vs CatBoost	0.058	[0.023, 0.095]	-	-
Mean predictor vs ICA-XGBoost	0.004	[−0.005, 0.016]	-	-
Mean predictor vs ICA-LightGBM	0.009	[−0.008, 0.027]	-	-
Mean predictor vs ICA-CatBoost	0.048	[0.013, 0.086]	-	-

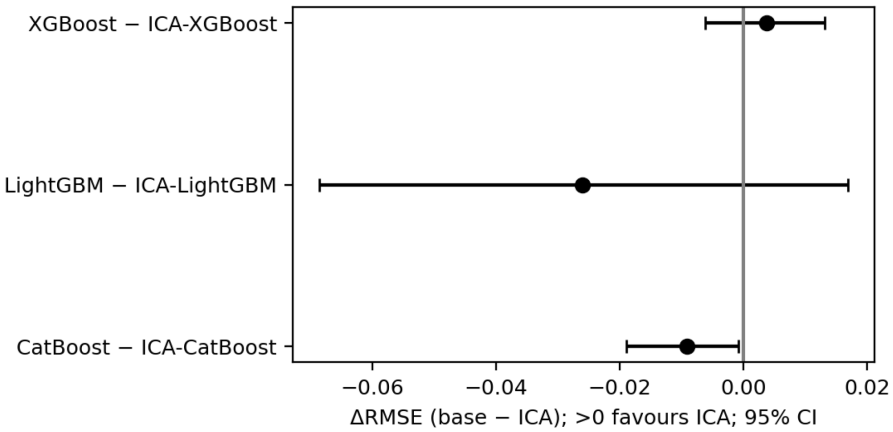


Figure 9. Paired RMSE difference, base learner minus ICA-tuned learner (positive favours ICA), with 95% family-bootstrap CIs.

Safety-critical performance is summarised in Figure 10 and Table 8 and Table 9.

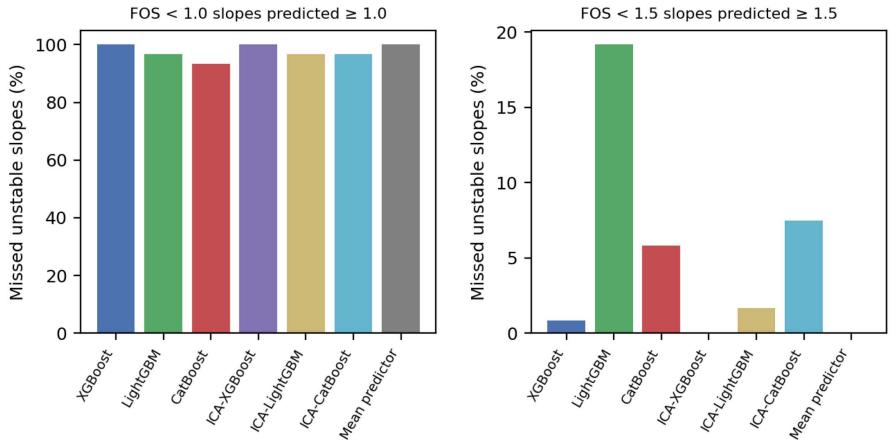


Figure 10. Share of slopes below each threshold that were predicted at or above it (missed unstable slopes), nested CV.

Table 8. Decision-threshold performance in nested CV (mean over 2 repeats; 15 of 80 development records have FOS < 1.0 and 60 have FOS < 1.5). "Missed" = true FOS below the threshold but predicted at or above it.

Model	FOS < 1.0 detected	Missed	Stable called unstable	Miss rate	FOS < 1.5 detected	Missed	Stable called unstable	Miss rate
XGBoost	0.0	15.0	0.0	100%	59.5	0.5	20.0	1%
LightGBM	0.5	14.5	4.0	97%	48.5	11.5	9.5	19%
CatBoost	1.0	14.0	1.5	93%	56.5	3.5	13.0	6%
ICA-XGBoost	0.0	15.0	0.0	100%	60.0	0.0	20.0	0%
ICA-LightGBM	0.5	14.5	1.0	97%	59.0	1.0	17.5	2%
ICA-CatBoost	0.5	14.5	2.5	97%	55.5	4.5	13.5	8%
Mean predictor	0.0	15.0	0.0	100%	60.0	0.0	20.0	0%

Table 9. Decision-threshold performance on the final held-out test set (n = 20; 4 records with FOS < 1.0, 15 with FOS < 1.5). Counts are very small; no precision should be inferred.

Model	FOS < 1.0 (4 unstable): missed	Stable called unstable	FOS < 1.5 (15 below): missed	Stable called unstable (of 5)
XGBoost	4	0	0	5
LightGBM	4	2	0	1
CatBoost	3	0	0	3
ICA-XGBoost	4	0	0	5
ICA-LightGBM	4	0	0	5
ICA-CatBoost	4	0	0	2
Mean predictor	4	0	0	5

In nested CV, only 15 of the 80 development records have FOS < 1.0; every model missed nearly all of them (predicting FOS ≥ 1.0), with miss rates of 93% (CatBoost) to 100% (XGBoost, ICA-XGBoost, and the mean-only reference). At the FOS = 1.5 threshold, XGBoost and ICA-XGBoost classified almost all slopes as below 1.5, which hides their inability to discriminate (all 20 stable records misclassified); LightGBM missed 19% of slopes below 1.5, and CatBoost and ICA-CatBoost 6% - 8%. On the final test set, 3 - 4 of the 4 unstable slopes were missed by every model. Aggregate regression metrics therefore overstate suitability for stability screening, and the models should not be used as a substitute for LEM analysis.

4.6. Implications for Model Selection

The evidence supports at most the following: CatBoost (regularised or ICA-tuned) provides small but statistically supported improvements over a mean-only predictor in group-aware nested CV, whereas XGBoost and LightGBM variants do not; ICA tuning with the overfitting penalty did not improve held-out error; and no model is accurate enough to classify unstable slopes. A model ranking scorecard was therefore omitted, because differences between most models are within sampling uncertainty.

4.7. Sensitivity Analysis

Figure 11 and Figure 12 present permutation importance (mean decrease in R^2 on outer test folds ± SD over 10 folds) for the best standalone model (CatBoost) and the best ICA hybrid (ICA-CatBoost); Table 10 gives all six models. Friction angle has the highest mean importance in all six models (CatBoost 0.27, ICA-CatBoost 0.30), followed by slope angle, slope height and cohesion, whose order differs between models and whose error bars overlap. For XGBoost, ICA-XGBoost and ICA-LightGBM, which show little or no skill, importance values are near zero and should not be interpreted. These importances describe the surrogate models of an LEM calculation; they are not independent evidence of field behaviour.

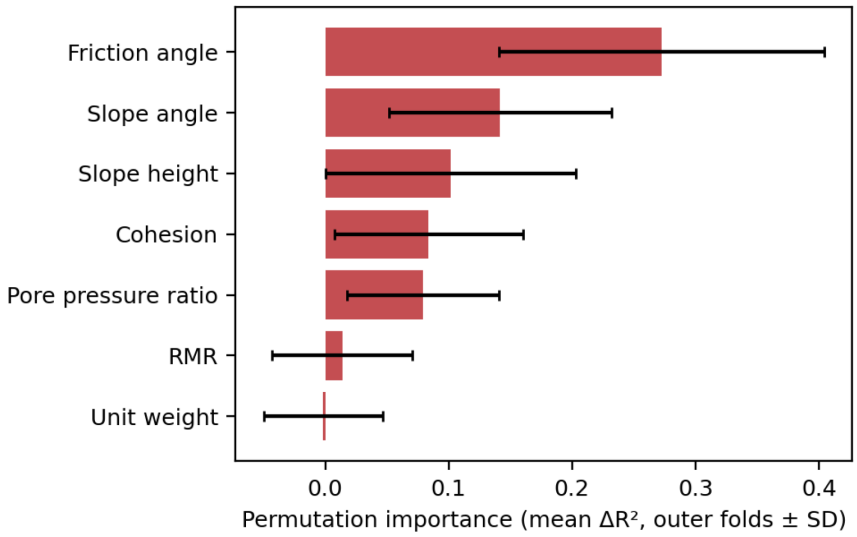


Figure 11. Permutation importance for regularised CatBoost (nested CV outer folds, mean ± SD).

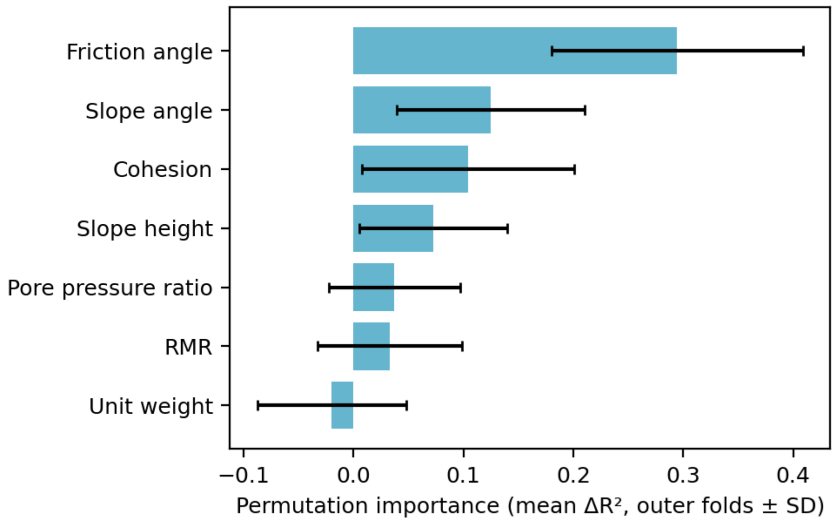


Figure 12. Permutation importance for ICA-CatBoost (nested CV outer folds, mean ± SD).

Table 10. Mean permutation importance (decrease in R^2) over 10 outer test folds.

Feature	XGBoost	LightGBM	CatBoost	ICA-XGBoost	ICA-LightGBM	ICA-CatBoost
RMR	−0.010	0.001	0.014	0.000	−0.001	0.033
Unit weight	−0.006	−0.017	−0.002	0.000	−0.000	−0.020
Cohesion	−0.000	0.186	0.084	0.004	0.015	0.105
Friction angle	0.066	0.562	0.272	0.019	0.113	0.295
Slope angle	0.022	0.188	0.142	0.009	0.070	0.125
Slope height	−0.009	0.132	0.102	0.000	0.021	0.073
Pore pressure ratio	−0.004	−0.008	0.079	−0.001	0.029	0.037

Because friction angle is the highest-ranked predictor in the surrogate models, characterising friction angle is a reasonable priority for LEM input quality in weathered units; this is a statement about the sensitivity of the modelled FOS to its inputs, not a validated site-risk finding.

4.8. Limitations

1) FOS is computed by Bishop's method from the same inputs used as predictors; the models are LEM surrogates and were not assessed against observed slope outcomes. 2) The analysis dataset has 100 unique records in 62 slope families; the final test set contains only 20 records (4 with FOS < 1.0), so estimates are imprecise. 3) Identical inputs with different FOS values show that some LEM inputs are not captured. 4) The compiled file has no per-record source identifier or Bishop-analysis settings, so source counts could not be reported and source-level generalisation could not be tested; slope families were used as a conservative grouping proxy. 5) Two repeats of 5-fold nested CV were used; the ICA budget (12 countries, 20 iterations) was fixed. 6) The $\lambda = 0.3$ coefficient was not tuned. 7) P-values are uncorrected and exploratory. Comparison with published studies is not made quantitatively, because published datasets, outputs and validation designs differ.

5. Conclusions

This study re-evaluated gradient boosting and ICA-tuned gradient boosting surrogates for Bishop-method FOS under duplicate-free, group-aware validation. The following conclusions are drawn.

1) Only 103 of 281 compiled records are unique (100 unique input vectors; 62 slope families). Random splitting of the raw data leaks information and inflates accuracy.

2) Under repeated group-aware nested CV, performance is low: CatBoost R^2 = 0.27 (95% CI [0.11, 0.40]), ICA-CatBoost R^2 = 0.23; other models are indistinguishable from a mean-only predictor. No "best model" can be claimed beyond CatBoost's consistent edge over XGBoost.

3) The ICA overfitting penalty was active but did not improve held-out error relative to an unpenalised search, and ICA tuning did not improve any base learner.

4) Friction angle is the highest-ranked predictor in all six surrogate models, followed by slope angle, height and cohesion with overlapping uncertainty; this concerns the sensitivity of the surrogate, not field behaviour.

5) The models missed almost all slopes with FOS < 1.0 and should not be used for stability screening. Larger, source-labelled datasets with independent outcome data and additional LEM inputs are needed.

Future work should assemble a larger, provenance-labelled dataset (with SHAP-based interpretation [26] where models show skill) with documented slip-surface and groundwater settings, evaluate against observed slope performance, and use group-aware validation as standard practice.

Acknowledgements

The authors gratefully acknowledge AngloGold Ashanti Iduapriem Mine, Tarkwa, for access to site geotechnical data and for technical discussions. The authors also acknowledge the University of Mines and Technology (UMaT) and the Festari Institute of Professional Studies, Tarkwa, for providing computational resources. This research received no external funding.

Data Availability Statement

The de-duplicated analysis dataset (100 records with slope-family identifiers and the development/test split), the validation procedure and the full analysis code are provided as Supplementary Material. The original 281-record compilation includes data obtained from third-party sources; these are available from the corresponding author with the permission of the data providers.

Author Contributions

Conceptualisation, F.K.-S.; Methodology, L.T.K. and F.K.-S.; Software, L.T.K.; Validation, L.T.K. and I.M.Q.; Formal analysis, L.T.K.; Investigation, I.M.Q. and F.K.-S.; Resources, F.K.-S.; Data curation, L.T.K. and I.M.Q.; Writing—original draft preparation, L.T.K. and C. A; Writing—review and editing, C. A.; Visualisation, L.T.K.; Supervision, F.K.-S. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Read, J. and Stacey, P. (2009) Guidelines for Open Pit Slope Design. CSIRO Publishing.

[2] Hoek, E. and Bray, J.W. (1981) Rock Slope Engineering. 3rd Edition, Institution of Mining and Metallurgy. https://www.researchgate.net/publication/313711304_Rock_slope_engineering?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoic2VhcmNoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19

[3] Duncan, J.M. (1996) State of the Art: Limit Equilibrium and Finite-Element Analysis of Slopes. *Journal of Geotechnical Engineering*, **122**, 577-596. https://doi.org/10.1061/(asce)0733-9410(1996)122:7(577)

[4] Akayuli, C., Ofosu, B., Nyako, S.O. and Opuni, K.O. (2013) The Influence of Observed Clay Content on Shear Strength and Compressibility of Residual Sandy Soils. *International Journal of Engineering Research and Applications*, **3**, 2538-2542. https://www.academia.edu/download/94455991/OO3425382542.pdf

[5] Kesse, G.O. (1985) The Mineral and Rock Resources of Ghana. A.A. Balkema. https://www.researchgate.net/publication/236368925_The_Mineral_and_Rock_Resources_of_Ghana?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoic2VhcmNoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19

[6] Bishop, A.W. (1955) The Use of the Slip Circle in the Stability Analysis of Slopes. *Géotechnique*, **5**, 7-17. https://doi.org/10.1680/geot.1955.5.1.7

[7] Morgenstern, N.R. and Price, V.E. (1965) The Analysis of the Stability of General Slip Surfaces. *Géotechnique*, **15**, 79-93. https://doi.org/10.1680/geot.1965.15.1.79

[8] Spencer, E. (1967) A Method of Analysis of the Stability of Embankments Assuming Parallel Inter-Slice Forces. *Géotechnique*, **17**, 11-26. https://doi.org/10.1680/geot.1967.17.1.11

[9] Fredlund, D.G. and Krahn, J. (1977) Comparison of Slope Stability Methods of Analysis. *Canadian Geotechnical Journal*, **14**, 429-439. https://doi.org/10.1139/t77-045

[10] Abramson, L.W., Lee, T.S., Sharma, S. and Boyce, G.M. (2002) Slope Stability and Stabilization Methods. 2nd Edition, John Wiley and Sons. https://www.researchgate.net/publication/255222923_Slope_Stability_and_Stabilization_Methods?_tp=eyJjb250ZXh0Ijp7Im-ZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InNlYXJjaCIs-InBvc2l0aW9uIjoicGFnZUhlYWRlciJ9fQ

[11] Chen, T. and Guestrin, C. (2016) XgBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3146-3154. https://www.researchgate.net/publication/378480234_LightGBM_A_Highly_Efficient_Gradient_Boosting_Decision_Tree?_tp=eyJjb250ZXh0Ijp7Im-ZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InNlYXJjaCIs-InBvc2l0aW9uIjoicGFnZUhlYWRlciJ9fQ

[13] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems*, **31**, 6638-6648. https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

[14] Atashpaz-Gargari, E. and Lucas, C. (2007) Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired by Imperialistic Competition. 2007 *IEEE Congress on Evolutionary Computation*, Singapore, 25-28 September 2007, 4661-4667. https://doi.org/10.1109/cec.2007.4425083

[15] Armaghani, D.J., Koopialipoor, M., Marto, A. and Yagiz, S. (2019) Application of Several Optimization Techniques for Estimating TBM Advance Rate in Granitic Rocks. *Journal of Rock Mechanics and Geotechnical Engineering*, **11**, 779-789. https://doi.org/10.1016/j.jrmge.2019.01.002

[16] Zhou, J., Zhu, S., Qiu, Y., Armaghani, D.J., Zhou, A. and Yong, W. (2022) Predicting Tunnel Squeezing Using Support Vector Machine Optimized by Whale Optimization Algorithm. *Acta Geotechnica*, **17**, 1343-1366. https://doi.org/10.1007/s11440-022-01450-7

[17] Al-Bared, M.A.M., Mustaffa, Z., Armaghani, D.J., Marto, A., Yunus, N.Z.M. and Hasanipanah, M. (2021) Application of Hybrid Intelligent Systems in Predicting the Unconfined Compressive Strength of Clay Material Mixed with Recycled Additive. *Transportation Geotechnics*, **30**, Article 100627. https://doi.org/10.1016/j.trgeo.2021.100627

[18] Goh, A.T.C. (1994) Seismic Liquefaction Potential Assessed by Neural Networks.

Journal of Geotechnical Engineering, **120**, 1467-1480. https://doi.org/10.1061/(asce)0733-9410(1994)120:9(1467)

[19] Sakellariou, M.G. and Ferentinou, M.D. (2005) A Study of Slope Stability Prediction Using Neural Networks. *Geotechnical and Geological Engineering*, **23**, 419-445. https://doi.org/10.1007/s10706-004-8680-5

[20] Choobbasti, A.J., Farrokhzad, F. and Barari, A. (2009) Prediction of Slope Stability Using Artificial Neural Network (Case Study: Noabad, Mazandaran, Iran). *Arabian Journal of Geosciences*, **2**, 311-319. https://doi.org/10.1007/s12517-009-0035-3

[21] Qi, C. and Tang, X. (2018) Slope Stability Prediction Using Integrated Metaheuristic and Machine Learning Approaches: A Comparative Study. *Computers & Industrial Engineering*, **118**, 112-122. https://doi.org/10.1016/j.cie.2018.02.028

[22] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[23] Kaveh, A. and Talatahari, S. (2010) Optimum Design of Skeletal Structures Using Imperialist Competitive Algorithm. *Computers & Structures*, **88**, 1220-1229. https://doi.org/10.1016/j.compstruc.2010.06.011

[24] Bieniawski, Z.T. (1989) Engineering Rock Mass Classifications. John Wiley and Sons. https://www.researchgate.net/publication/236562758_Engineering_Rock_Mass_Classifica-tion?_tp=eyJjb250ZXh0Ijp7Im-ZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoic2Vhcm-NoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19

[25] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[26] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. 31*st Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4765-4774. https://doi.org/10.48550/arXiv.1705.07874