Green Energy and Resources 4 (2026) 100203

Contents lists available at ScienceDirect

Green Energy and Resources

journal homepage: www.journals.elsevier.com/green-energy-and-resources

Research Article

Rock mechanics–driven interpretable machine learning framework for preliminary seal integrity screening in geological CO_2 storage

Muhammad Hammad Rasool*, A.K.M. Eahsanul Haque*, Numair Ahmed Siddiqui, Hassan Soleimani, Maqsood Ahmad*, Nisar Ahmed, Muhammad Mubashir, Venkateshwaran Baskaran, Md Nasirul Islam, Lawal Opeyemi Olajid

Geosciences Department, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Malaysia

ARTICLE INFO

Keywords:
Caprock integrity
CO_2 geological storage
Stress-controlled behaviour
Interpretable machine learning
Random forest
Decision tree

ABSTRACT

Seal integrity assessment is essential for ensuring the long-term security of geological CO_2 storage; however, early-stage evaluation is often challenged by limited and heterogeneous subsurface datasets. This study proposes an interpretable machine learning framework for preliminary seal integrity screening using limited rock mechanics data from the 31/5–7 Eos well-core of the Northern Lights CO_2 storage project. An unsupervised learning workflow was first employed to reconstruct Safe, Transitional, and Risk integrity regimes through cluster validation and ambiguity-based boundary analysis. Multiple supervised learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost, were subsequently developed and evaluated under stratified train/test split, Leave-One-Out Cross Validation (LOOCV), and repeated stratified cross-validation strategies. Random Forest demonstrated the most stable predictive behaviour across validation schemes, achieving balanced accuracies exceeding 0.86. Explainability analyses using Random Forest feature importance and SHAP consistently identified porosity and failure related effective axial stress as dominant screening controls, suggesting coupled stress–structure interactions govern integrity behaviour within the studied dataset. Based on these observations, a Stress–Porosity Screening Score (SPSS) was developed as a simplified post hoc screening descriptor for rapid integrity evaluation. Statistical analyses demonstrated significant differentiation of SPSS across reconstructed integrity regimes ($p < 0.001$). Decision-tree interpretation further enabled extraction of preliminary data-emergent screening pathways and threshold behaviour. Conclusively, the proposed methodology provides an interpretable and data-adaptive screening approach intended to support early-stage caprock assessment and guide subsequent detailed geomechanical investigations.

1. Introduction

The long-term success of geological carbon storage depends fundamentally on the integrity of caprock formations, which act as the primary barrier preventing upward migration of injected CO_2 (Ahmed et al., 2025; Rasool et al., 2023). Any compromise in caprock stability whether through mechanical failure, fracture propagation, or the creation of leakage pathways can significantly undermine storage security (Ross and Alexander, 2026). As large-scale carbon capture and storage (CCS) projects continue to expand globally, reliable assessment of caprock integrity under in-situ stress conditions has become a central requirement for safe and sustainable subsurface operations (Younessi et al., 2024). This need is particularly critical during early-stage site screening, where decisions must often be made with limited and heterogeneous data (Hannon and Esposito, 2015).

Conventional approaches to caprock integrity evaluation are predominantly based on analytical and semi-empirical geomechanical models, such as Mohr–Coulomb (Lee et al., 2023) and Hoek–Brown (Rafiei Renani and Cai, 2022) failure criteria. These frameworks provide a physically grounded basis for understanding rock failure; however, they rely on assumptions of material homogeneity, well-constrained laboratory conditions, and consistent datasets (Suhaidi et al., 2025). In practice, subsurface datasets are inherently heterogeneous, often compiled from multiple lithologies, testing methods, and stress conditions (Kim et al., 2025). This variability limits the direct applicability of classical models and introduces uncertainty in predicting caprock

* Corresponding authors.
E-mail addresses: Muhammad_19000949@utp.edu.my, hammad.rasool@utp.edu.my (M.H. Rasool), eahsanul.haque@utp.edu.my (A.K.M.E. Haque), Maqsood.ahmad@utp.edu.my (M. Ahmad).

https://doi.org/10.1016/j.gerr.2026.100203
Received 27 April 2026; Received in revised form 15 June 2026; Accepted 29 June 2026
2949-7205/© 2026 The Authors. Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

behaviour at field scale. As a result, there is a growing recognition that traditional approaches, while essential, may not fully capture the complexity of real-world CCS systems (Venkateshwaran et al., 2023).

Recent advances in data-driven techniques, particularly machine learning (ML), offer new opportunities to analyze such heterogeneous datasets (Isah et al., 2025). ML methods can identify patterns and nonlinear relationships without requiring strict assumptions about material behaviour, making them attractive for subsurface applications (Rasool et al., 2025). Recent studies have increasingly explored the application of machine learning techniques for evaluating caprock integrity and related geomechanical properties in CO_2 storage systems. For instance, the work by Yao et al. (2023) provide a comprehensive overview of ML applications in CCS, highlighting the use of artificial neural networks (ANN), support vector machines (SVM), and ensemble methods such as Random Forest (RF) for predicting rock properties, deformation behaviour, and leakage risk. Similarly, Aminaho et al. (2025) developed ANN-based models to predict rock brittleness for CO_2 geosequestration, demonstrating that ML can effectively approximate complex geochemical–geomechanical relationships; however, their approach relied heavily on numerically simulated datasets rather than experimentally validated field data, limiting its direct applicability to real subsurface conditions. Additionally, studies such as Tillero (2024) employed ML to accelerate CO_2 storage site appraisal by predicting trapping efficiency and pressure behaviour, yet these models primarily focus on reservoir-scale performance rather than caprock failure mechanisms.

Furthermore, recent research has also highlighted the critical role of geomechanical, hydromechanical, and petrophysical processes in governing caprock integrity during CO_2 storage. Studies by Skerbisch et al. (2025), Hawez et al. (2025), Sadiq et al. (2025), and Suhaidi et al. (2025) identified stress evolution, permeability changes, fault reactivation, and monitoring requirements as major factors influencing seal performance. More recently, de Lima et al. (2026) demonstrated the growing potential of machine learning for predicting fault reactivation and assessing storage integrity risks.

Previous studies mostly emphasize predictive performance rather than physical interpretability, often treating caprock integrity as a purely data-driven classification problem. This creates a disconnect between model outputs and engineering decision-making, where transparent and physically meaningful indicators are required. Consequently, the challenge is not only to improve predictive accuracy, but to develop approaches that integrate data-driven insights with fundamental geomechanical understanding (Sun et al., 2023). Another limitation in current caprock assessment frameworks is the absence of simple, robust screening parameters that can be applied consistently across datasets with varying quality and completeness (Karpfinger et al., 2025). Caprock integrity is influenced by both mechanical stability and transport properties, yet integrating these effects into a unified and practical metric remains difficult. In many datasets, permeability measurements are sparse or exhibit limited variability, while stress-related parameters dominate the available information. This imbalance complicates the development of coupled hydro-mechanical models and highlights the need for simplified approaches that remain physically meaningful under data constraints.

This study addresses these limitations by developing an interpretable machine learning framework for preliminary seal integrity screening using lab measured rock mechanics data from the 31/5–7 Eos well-core of the Northern Lights CO_2 storage system. The proposed workflow integrates unsupervised regime reconstruction, multiple supervised learning models, explainable AI techniques, and a stress–porosity screening concept to identify dominant controls governing integrity behaviour and establish preliminary screening pathways. This work emphasizes physically interpretable decision support by linking machine learning outputs with rock mechanics behaviour and uncertainty-aware screening domains. The framework is intended as a preliminary assessment tool capable of supporting early-stage site evaluation and guiding subsequent detailed geomechanical investigations.

2. Methodology

2.1. Data source and geological context

The dataset used in this study consists of rock mechanics data from the 31/5–7 Eos well core developed for the Northern Lights CCS project. This dataset represents a comprehensive experimental rock mechanics testing program conducted by the Norwegian Geotechnical Institute (NGI) for Equinor, aimed at characterizing both reservoir and caprock formations relevant to CO_2 storage in the North Sea.

The dataset comprises a comprehensive (but limited 43 samples) rock mechanics testing program, including triaxial compression, hydrostatic, and uniaxial strain tests to characterize strength and deformation under in-situ stress conditions. Brazilian tensile strength tests were conducted to evaluate tensile failure behaviour, alongside permeability measurements on both vertical and horizontal core plugs to capture transport properties and anisotropy. These combined experiments provide an integrated dataset representing both mechanical stability and fluid flow behaviour relevant to caprock integrity assessment.

The data originate from the 31/5–7 (Eos) exploration well located in the Aurora license within the Horda Platform region, a key site selected for large-scale geological CO_2 storage. The well penetrates the Lower Jurassic Dunlin Group storage complex, where the Johansen and Cook formations serve as reservoir units, while the overlying Drake Formation specifically the Intra-Drake shale interval acts as the primary sealing caprock. Core samples used in this study span a depth interval of approximately 2592 to 2761 m, encompassing both sandstone and shale lithologies. Specifically, sandstone samples were obtained from the Cook Formation (2680.25–2680.5 m), Johansen 3 Formation (~2725 m), and Johansen 4 Formation (2761.00–2761.25 m), while shale samples were derived from the Intra-Drake Formation (2592.00–2595.18 m). These formations collectively represent a reservoir–caprock system as shown in Fig. 1 and Table 1.

It is important to point out here that the inclusion of both sandstone and shale samples reflects the integrated reservoir–seal architecture of the Northern Lights system rather than an attempt to model caprock behaviour using shale alone. During early-stage screening, subsurface datasets are often heterogeneous and compiled from multiple lithologies. Therefore, lithology was retained as a categorical descriptor to preserve geological context while allowing the framework to evaluate relative integrity tendencies across the available rock mechanics dataset. The reconstructed screening regimes should consequently be interpreted as data-derived domains rather than lithology-specific failure states.

2.2. Unsupervised regime construction and label reconstruction

An unsupervised learning workflow was implemented to construct preliminary caprock integrity screening regimes from the processed geomechanical and petrophysical dataset. Prior to clustering, numerical variables were standardized using StandardScaler to minimize scale-induced bias during distance-based learning. K-means clustering was initially evaluated using multiple cluster configurations ($k = 2$–5) to identify appropriate integrity group structures. Cluster quality was assessed using Silhouette score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI), together with cluster-size evaluation to identify instability and singleton behaviour.

Although three integrity categories (Safe, Transitional, and Risk) were initially targeted, the $k = 3$ solution produced highly imbalanced clusters. Consequently, a stable two-cluster configuration was first established, after which ambiguity-based boundary analysis was introduced to reconstruct intermediate conditions. Transitional samples were identified through uncertainty analysis near cluster boundaries to better

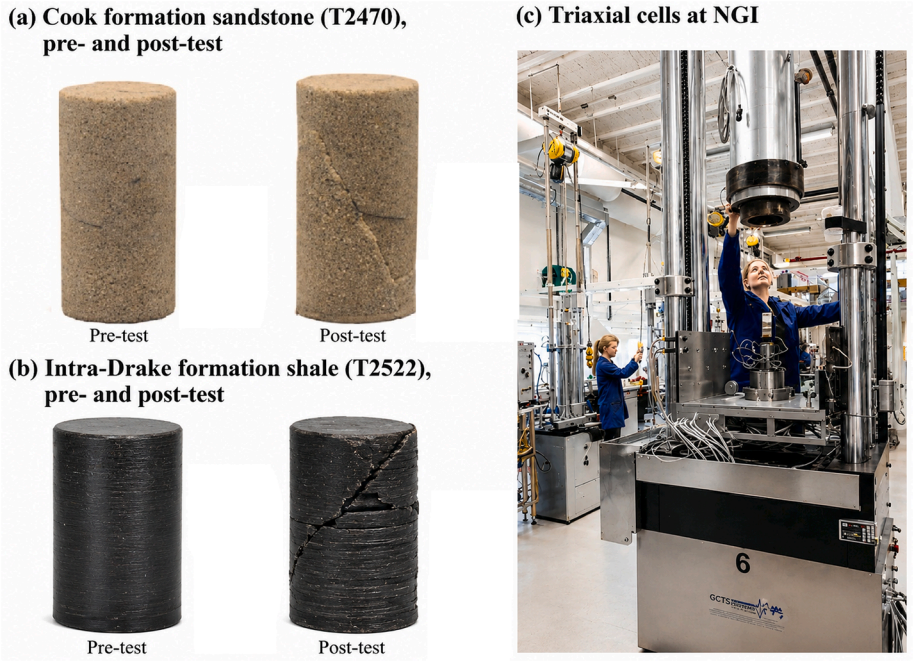


Fig. 1. Representative core samples and experimental setup used in the rock mechanics testing program. (a) Cook Formation sandstone (T2470) before and after triaxial testing, showing visible fracture development under applied stress. (b) Intra-Drake Formation shale (T2522) pre- and post-test, illustrating deformation and failure characteristics typical of low-permeability caprock materials. (c) Triaxial testing system at the Norwegian Geotechnical Institute (NGI), used to simulate in-situ stress conditions and evaluate mechanical behaviour under controlled laboratory environments (Griffiths et al., 2025).

Table 1
Summary of geomechanical, petrophysical, geological, and reconstructed variables used within the caprock integrity screening workflow. Depth was evaluated during sensitivity analysis but excluded from the final model to reduce potential proxy dominance and label leakage effects.

Nomenclature in this study	Category	Unit	Original dataset nomenclature	Description
Confining stress	Mechanical	MPa	confining_pressure_MPa	Applied confining pressure during testing
Peak stress	Mechanical	MPa	shear_stress_MPa_peak	Peak shear stress measured during triaxial compression testing
Axial stress	Mechanical	MPa	eff_vertical_stress_MPa_failure	Effective vertical stress recorded at specimen failure
Porosity	Petrophysical	%	porosity_pct_initial	Initial porosity before testing
Permeability	Petrophysical	mD	permeability_mD	Matrix permeability of shale
Lithology	Geological	–	material_lithology	Sandstone and Shale
Depth	Geological	m	depth_m	2592m-2671m
Integrity regimes	Clustering derived	–	Safe/Transitional/Risk	Manually labelled based on geomechanics
SPSS	Derived	–	Dataset dependent	Derived based on SHAP results

represent gradual integrity evolution. In addition, sensitivity analysis was performed to evaluate potential dominance of proxy variables, particularly depth, resulting in reconstruction of the final depth-independent Safe–Transitional–Risk screening labels. The engineering labels (Safe, Transitional, and Risk) were assigned to the reconstructed clusters through expert geomechanical interpretation.

2.3. Machine learning development, hyperparameter optimization and validation

Supervised machine learning models were subsequently developed to classify the reconstructed integrity regimes. Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) algorithms were evaluated owing to their differing assumptions regarding linearity and nonlinear feature interactions.

Hyperparameter optimization was performed using GridSearchCV coupled with Stratified K-Fold cross-validation ($n = 3$). Balanced accuracy was selected as the optimization metric because of moderate class imbalance within the reconstructed integrity labels. Model-specific parameter grids included optimization of regularization parameters for Logistic Regression, tree depth and splitting criteria for Random Forest and Decision Tree frameworks, kernel and regularization settings for SVM, and estimator configuration for XGBoost.

To evaluate predictive robustness, three complementary validation strategies were implemented: (i) stratified train/test split (Bhagat and Bakariya, 2022), (ii) Leave-One-Out Cross Validation (LOOCV) (Khadka, 2023), and (iii) Repeated Stratified K-Fold validation (Widodo et al., 2022). Performance was assessed using accuracy, balanced accuracy, and Macro F1-score. Balanced accuracy and Macro F1 metrics were additionally emphasized because they provide class-sensitive evaluation under non-uniform class distributions. A stratified 70:30 train–test split was employed to preserve class proportions across reconstructed integrity regimes. Owing to the limited sample size, model performance was evaluated using LOOCV and repeated stratified K-fold cross-validation to reduce partition sensitivity and improve reliability assessment.

2.4. Model interpretability and threshold extraction

Following model evaluation, explainability analyses were conducted to identify dominant controls governing reconstructed integrity regimes. Random Forest was selected as the primary interpretation framework due to its comparatively stable performance across validation strategies. Variable influence was first evaluated using Random Forest feature importance, followed by SHapley Additive exPlanations (SHAP) to quantify global and class-specific feature contributions.

To improve interpretability and convert learned relationships into transparent screening pathways, a Decision Tree framework was additionally employed for extraction of data-emergent screening boundaries. The Decision Tree was used solely as an interpretability tool rather than for predictive optimization. Extracted thresholds were interpreted as

preliminary screening indicators specific to the studied dataset rather than universal geomechanical limits.

2.5. Stress–Porosity Screening Score (SPSS)

The Stress–Porosity Screening Score (SPSS) was developed as a post hoc screening descriptor following explainability analysis. SPSS was motivated by the consistent emergence of porosity and axial stress as dominant predictors within the reconstructed screening framework across Random Forest and SHAP analyses. SPSS was introduced as a simplified data-adaptive descriptor intended to summarize the principal variables governing integrity behaviour.

Axial stress and porosity were normalized using Min–Max scaling and combined through weights derived from normalized SHAP contributions as shown in eq. 1

$$SPSS = w_1 A_n + w_2 P_n \ldots \quad \text{eq. (1)}$$

where A_n and P_n represent normalized axial stress and porosity, respectively, while w_1 and w_2 denote SHAP-derived normalized feature weights. Statistical significance of SPSS across reconstructed screening regimes was evaluated using ANOVA and Kruskal–Wallis testing.

2.6. Seal integrity framework design

The collective samples were initially preprocessed and normalized to facilitate the clustering task. Subsequently, each cluster was manually assessed through ambiguity analysis and regime classification to establish reliable labels. A depth influence assessment was also incorporated to verify adequate depth coverage and sample distribution. The prepared and calibrated dataset was then utilized to train the selected machine learning models for the classification tasks. Furthermore, a novel metric, the Stress–Porosity Screening Score (SPSS), was introduced to explain the classification performance.

The overall workflow is shown in Fig. 2.

3. Results and discussion

3.1. Construction of preliminary integrity screening regimes

Initial K-means clustering and cluster validation analyses were performed using multiple cluster configurations ($k = 2$–5) together with Silhouette, Davies–Bouldin, and Calinski–Harabasz metrics. Although three classes were initially targeted to represent Safe, Transitional, and Risk conditions, the $k = 3$ solution produced highly imbalanced clusters,

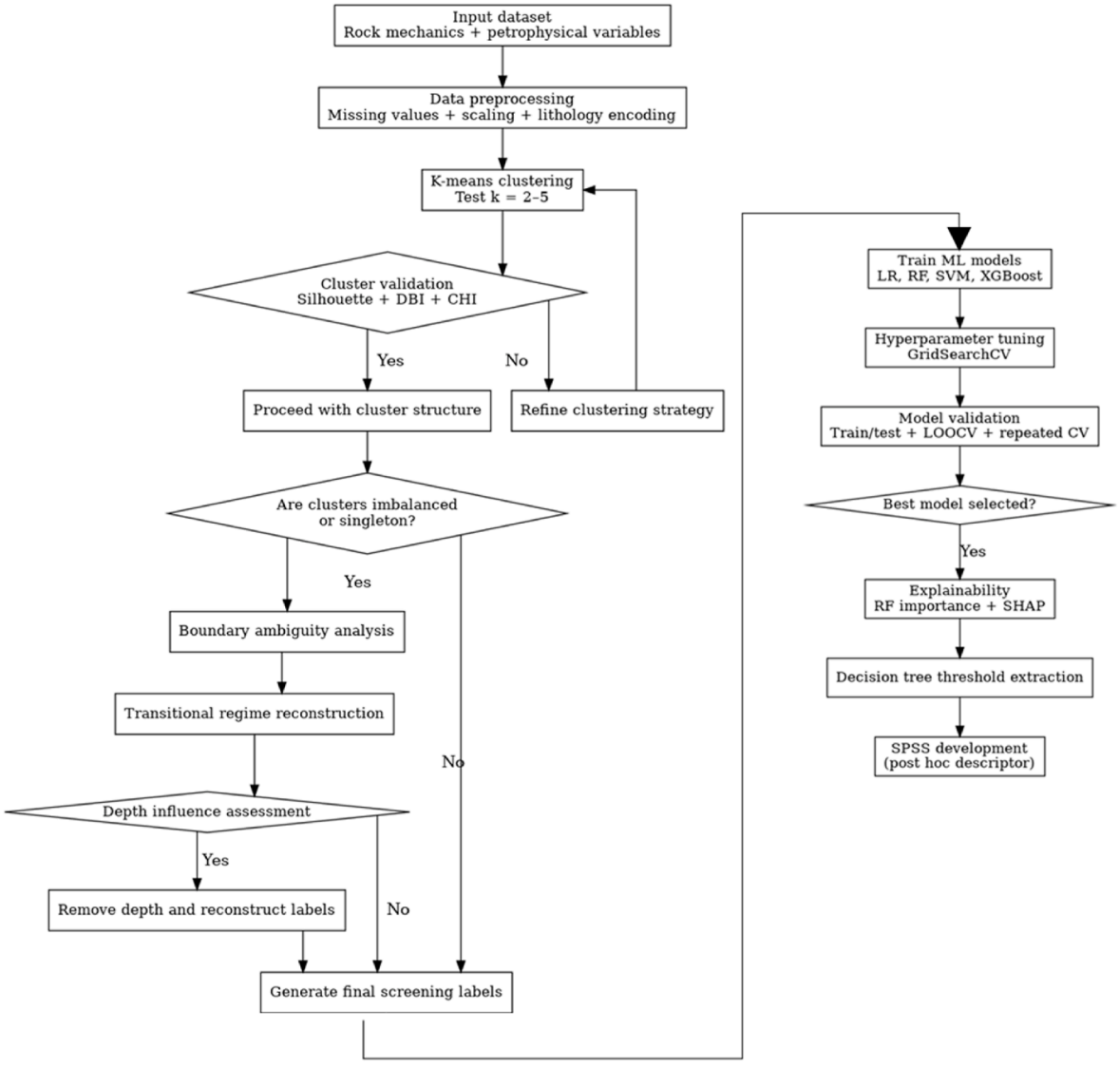


Fig. 2. Overall workflow of the research.

including singleton behaviour. Consequently, a stable two-cluster structure (Safe and Risk) was first established, after which ambiguity-based boundary analysis was implemented to reconstruct the Transitional regime and better capture intermediate integrity behaviour. Detailed ambiguity based boundary analysis can be found in the colab notebook in code availability section.

The finalized integrity regimes exhibited partially separable but non-discrete distributions following exclusion of depth effects, indicating the presence of both distinct and transitional integrity behaviour. Safe samples occupied relatively compact regions within the reduced feature space, suggesting greater consistency in their geomechanical and petrophysical characteristics. In contrast, Risk observations displayed broader dispersion, while Transitional samples predominantly occupied intermediate regions between Safe and Risk domains rather than forming isolated populations. Similar trends were observed in both PCA and t-SNE projections (Figs. 3 and 4).

The nonlinear t-SNE embedding further revealed structural complexity not fully captured by PCA. Partial overlap between Risk and Transitional observations suggests that elevated integrity conditions may arise through multiple combinations of controlling variables rather than a single deterministic pathway. The observed continuity across classes indicates gradual integrity evolution and supports the implementation of ambiguity-based calibration and uncertainty-aware reconstruction rather than direct hard-cluster assignment.

Overall, the reconstructed Safe–Transitional–Risk labels represent integrity screening regimes derived from data-driven clustering and geomechanical interpretation rather than deterministic geological states. The observed overlap and transitional structure justify the subsequent supervised learning framework for capturing complex integrity behaviour.

3.2. *Performance optimization and validation of machine learning screening models*

The machine learning models demonstrated varying predictive capability across validation strategies despite undergoing hyperparameter optimization using GridSearchCV. Random Forest consistently exhibited the strongest performance, achieving accuracies of 84.6%, 86.1%, and 82.9% under stratified train/test, LOOCV, and repeated stratified cross-validation, respectively, and its confusion matrix are shown in Fig. 5. The moderate prediction performance of Random Forest should be interpreted within the context of the limited dataset size and reconstructed integrity labels, where restricted sample diversity and class representation may constrain the model's ability to fully capture complex subsurface behaviour. Similar trends were observed for balanced accuracy and Macro F1-score, indicating relatively stable behaviour across validation schemes. The limited variation in performance suggests moderate prediction consistency despite the relatively small sample size and reconstructed nature of the screening labels (Table 2).

For briefness, only the confusion matrix corresponding to the best-performing model (Random Forest) is presented in the main manuscript. Complete confusion matrix, additional robustness analyses, and model evaluation outputs for all algorithms and validation strategies are provided within the supporting Google Colab notebook with python script in code availability section.

Distinct differences were observed between linear and nonlinear approaches. Logistic Regression exhibited the weakest performance, particularly under LOOCV (46.5%), whereas Random Forest, SVM, and XGBoost consistently achieved higher predictive capability as shown in Fig. 6. The lower performance of Logistic Regression suggests that the reconstructed integrity regimes may involve nonlinear interactions among geomechanical and petrophysical variables rather than simple linear relationships. The stronger behaviour of ensemble-based approaches further supports the existence of complex feature interactions influencing integrity screening behaviour.

Interestingly, XGBoost exhibited noticeable improvement under repeated cross-validation compared with the single train/test split, suggesting sensitivity to partitioning effects within the limited dataset. In contrast, Random Forest maintained comparatively stable behaviour across all validation strategies, indicating reduced susceptibility to data partition variability. Consequently, Random Forest was selected as the primary model for subsequent explainability analysis and identification

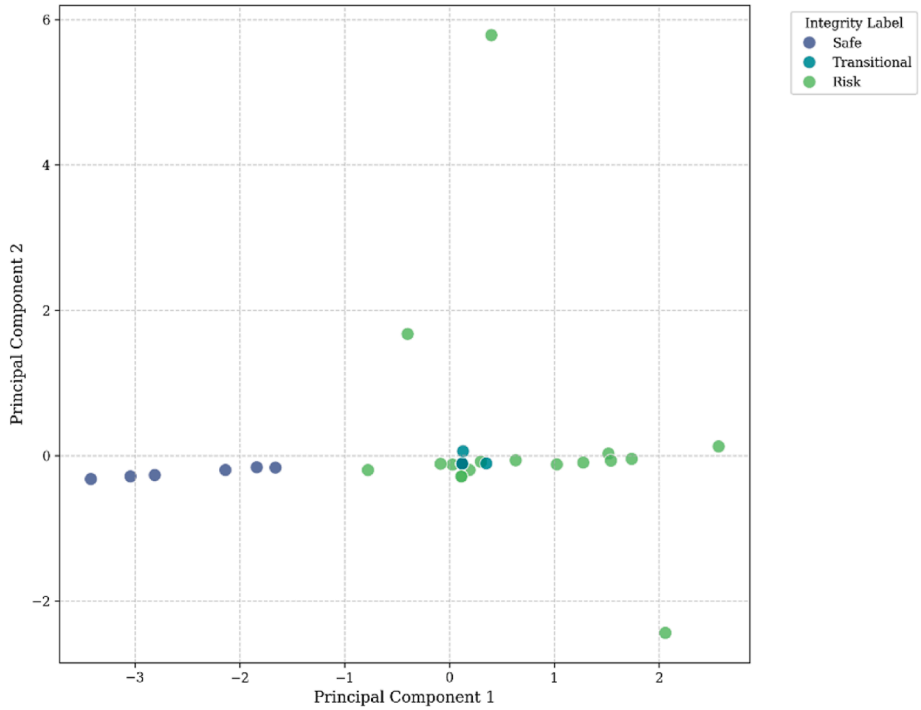


Fig. 3. Principal Component Analysis (PCA) projection of the reconstructed Safe, Transitional, and Risk screening regimes. Partial separation with intermediate overlap suggests gradual integrity evolution rather than distinct class boundaries.

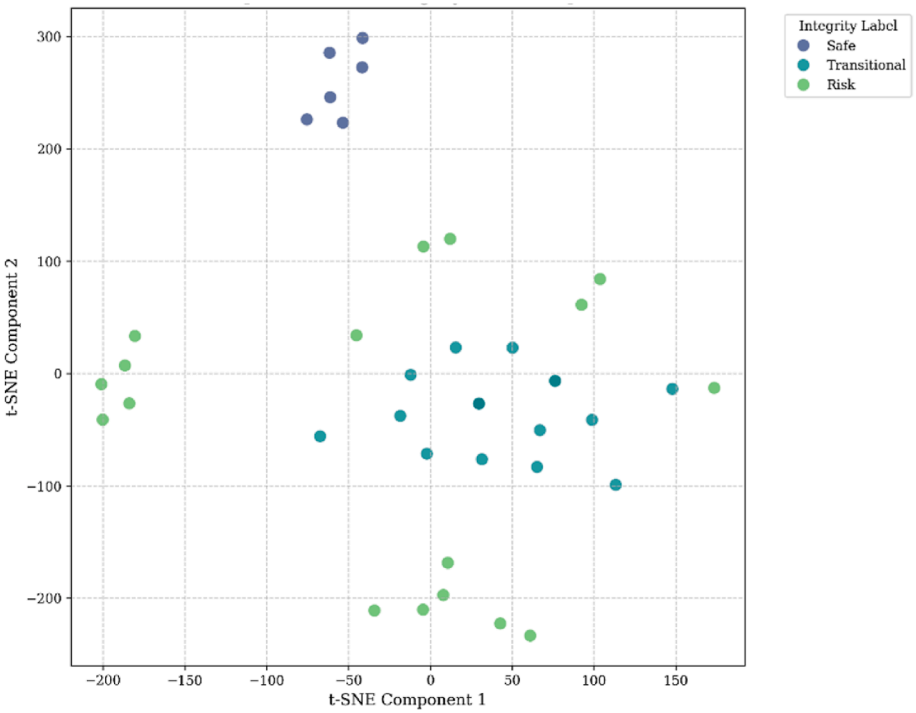


Fig. 4. t-SNE visualization of the reconstructed integrity screening regimes highlighting nonlinear relationships within the dataset. The overlap between Transitional and Risk samples indicates uncertainty near regime boundaries and supports ambiguity-based reconstruction.

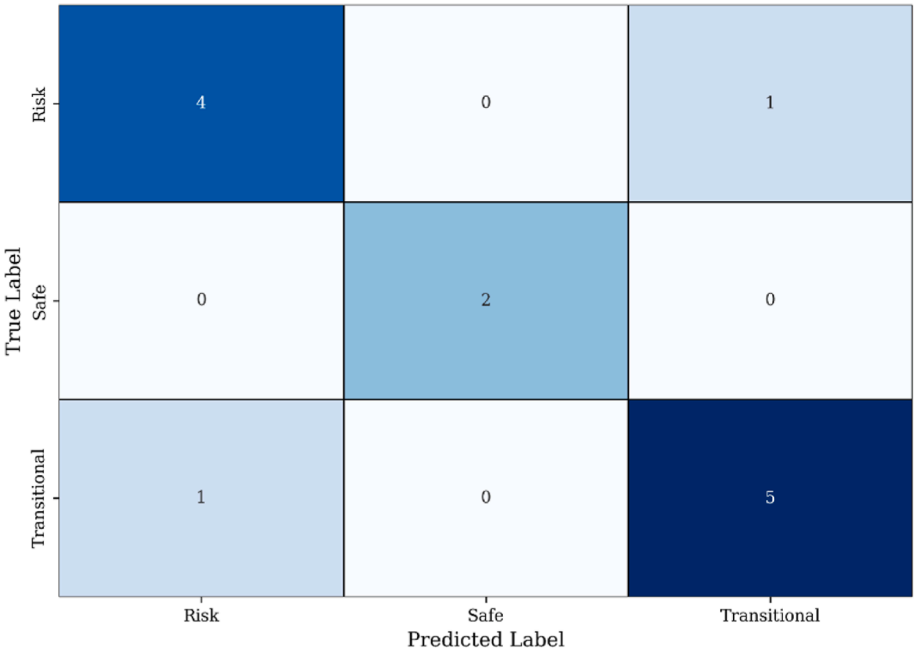


Fig. 5. Confusion matrix of the optimized Random Forest model under stratified train/test evaluation. The model demonstrated strong classification capability with minimal misclassification, particularly for Safe and Transitional regimes, indicating effective discrimination of reconstructed integrity screening labels.

of dominant screening controls. However, these results should be interpreted as preliminary screening performance rather than generalized predictive capability owing to sample size limitations and the reconstructed nature of the integrity labels.

3.2.1. Interpretable decision boundaries and threshold extraction

Although Random Forest demonstrated the strongest predictive performance and was selected as the primary screening model, a Decision Tree framework was additionally employed as an interpretability tool to extract preliminary data-emergent screening boundaries as illustrated in Fig. 7. The purpose of this analysis was to convert learned relationships into transparent decision pathways capable of supporting preliminary screening workflows. The resulting tree indicated that porosity governed the first and most influential split, with a threshold value of approximately −1.094 (standardized value) separating lower-risk conditions from more complex integrity states.

Table 2
Performance of optimized machine learning models under multiple validation strategies following hyperparameter tuning using GridSearchCV. Accuracy, balanced accuracy, and Macro F1-score are reported to evaluate predictive robustness and class-wise screening performance across reconstructed integrity regimes.

Model	Hyperparameter tuning strategy	Stratified train/test accuracy	LOOCV accuracy	Repeated stratified K-fold accuracy	Repeated CV balanced accuracy	Repeated CV macro F1
Logistic regression	GridSearchCV (C = 0.01–10)	0.615	0.465	0.605	0.681	0.670
Decision tree	GridSearchCV (max_depth, min_samples_split)	0.769	0.814	0.760	0.803	0.776
Random forest	GridSearchCV (n_estimators, max_depth)	0.846	0.861	0.830	0.861	0.863
SVM	GridSearchCV (kernel, C)	0.769	0.837	0.752	0.794	0.787
XGBoost	GridSearchCV (n_estimators, max_depth, learning rate)	0.692	0.791	0.814	0.849	0.831

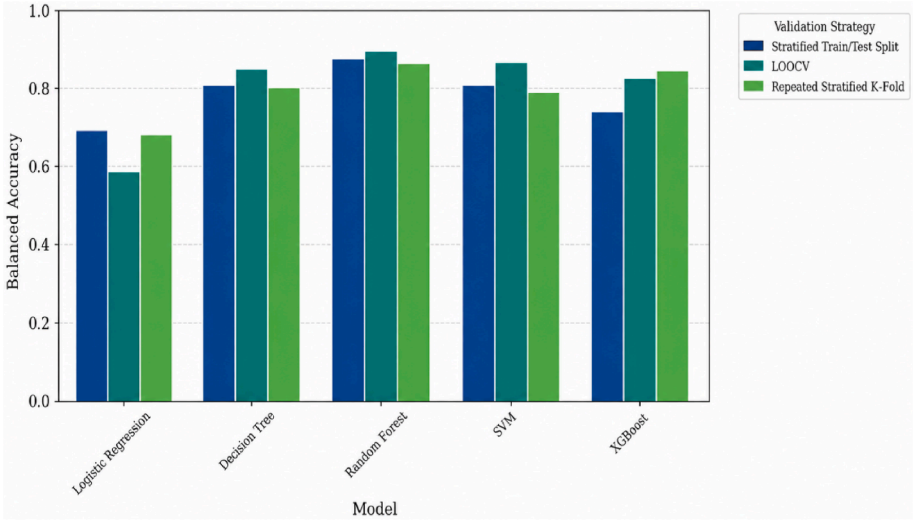


Fig. 6. Comparison of model performance across stratified train/test split, LOOCV, and repeated stratified K-fold validation strategies using balanced accuracy as the evaluation metric. Random Forest exhibited the highest and most consistent performance across validation schemes, while nonlinear approaches generally outperformed Logistic Regression.

Samples below this threshold transitioned directly toward the Safe regime, whereas samples exceeding this value required further discrimination through stress-related variables. This initial separation supports previous explainability analyses in which porosity consistently emerged as a dominant predictor of integrity behaviour.

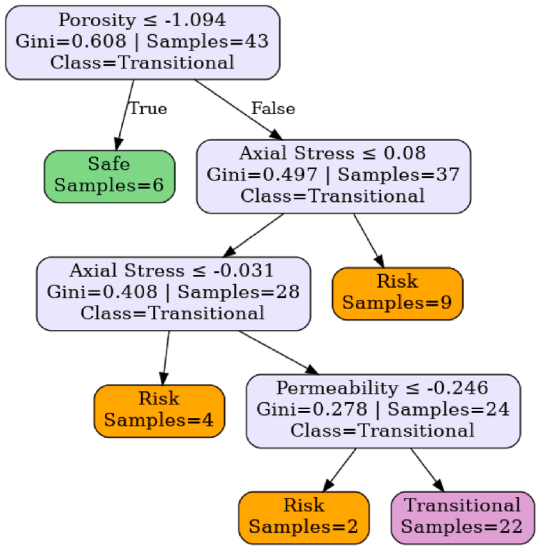


Fig. 7. Interpretable Decision Tree used for extraction of preliminary standardized data-emergent screening boundaries. The tree highlights porosity, axial stress (at failure), and permeability as hierarchical splitting variables governing transitions among Safe, Transitional, and Risk integrity regimes.

Further branching was primarily governed by axial stress thresholds of approximately 0.08 and −0.031 (standardized values), followed by a secondary permeability threshold near −0.246. Higher axial stress values directed samples toward Risk classifications, whereas intermediate stress conditions retained Transitional behaviour. This hierarchical progression suggests that integrity evolution is not controlled by a single geomechanical factor but by sequential interactions between pore structure and stress-state characteristics. Importantly, these thresholds should not be interpreted as universal geomechanical limits because they were extracted from a reconstructed dataset with limited sample size and standardized feature scaling. Instead, they represent preliminary screening indicators that provide interpretable support to the broader machine learning framework and help bridge predictive modelling with physically understandable decision boundaries.

3.3. *Explainable identification of dominant predictors*

The dominant predictors within the reconstructed screening framework identified through both Random Forest feature importance and SHAP interpretation demonstrated strong consistency, suggesting that the observed feature rankings were not artifacts of a single analytical method. Porosity and axial stress emerged as the primary contributors, followed by shale lithology, whereas permeability and peak stress exhibited comparatively weaker influence. Confining stress showed

negligible contribution across both analyses, indicating limited discriminatory capability within the present dataset (Figs. 8 and 9). Fig. 9 presents the SHAP-based interpretation of the Random Forest model used for seal integrity classification. As shown in Fig. 9(a), porosity and axial stress are the most influential predictors, with porosity exhibiting the highest contribution to the Safe and Transitional classes, while '*lithology shale*' and '*axial stress*' strongly influence the Risk class. The SHAP summary plot (Fig. 9(b)) further confirms the dominant role of these variables by displaying the largest spread of SHAP values across samples. Consistently, the heatmap in Fig. 9(c) highlights porosity, axial stress, and lithological characteristics as the primary controls governing caprock integrity predictions, whereas confining stress contributes negligibly to model performance. The negligible importance of confining stress likely reflects the limited range and distribution of the available dataset rather than a lack of physical relevance. The agreement between model-derived importance and explainability-based interpretation increases confidence that the identified variables reflect meaningful screening controls rather than algorithm-specific behaviour.

The dominance of porosity is physically plausible because pore structure directly influences rock framework characteristics, including grain contact density, compressibility, and deformation susceptibility. Variations in porosity can alter stress redistribution pathways and modify the mechanical response of the rock matrix under loading conditions. Samples with differing pore architectures may therefore exhibit substantially different integrity behaviour despite having similar stress conditions. The strong contribution of axial stress further suggests that externally applied loading conditions remain important drivers of integrity response, indicating that screening behaviour likely arises through coupled stress–structure interactions rather than through isolated geomechanical parameters alone.

Lithological variables, particularly shale occurrence, demonstrated moderate influence but greater influence than 'sandstone', suggesting that mineralogical composition contributes to integrity behaviour but does not independently dominate the reconstructed screening regimes. In contrast, permeability and peak stress contributed relatively weak influence, while confining stress exhibited near-zero importance because of the limited range and distribution of the available dataset rather than a lack of physical relevance. Therefore, the identified dominant controls should be interpreted as data-derived screening behaviour specific to the studied conditions rather than universal indicators governing caprock integrity across all geological systems.

3.4. Development of Stress–Porosity Screening Score (SPSS)

The dominance of porosity and axial stress across both Random Forest and SHAP analyses motivated the development of the Stress–Porosity Screening Score (SPSS) as a post hoc screening descriptor as shown in Fig. 10. The SPSS was designed to summarize the principal controls consistently identified by explainable machine learning. The rationale is physically meaningful because porosity governs pore architecture, grain-contact density, and deformation susceptibility, whereas axial stress (at failure) reflects the rock's loading state under externally imposed loading conditions. Their combined influence suggests that integrity behaviour may arise through coupled stress–-structure interactions rather than isolated geomechanical variables.

The proposed SPSS may offer practical utility as a rapid preliminary indicator for seal integrity evaluation during early-stage CO_2 storage assessment. By integrating the dominant controls identified through explainable machine learning into a simplified descriptor, SPSS may assist in identifying intervals associated with elevated integrity concern and prioritizing zones for detailed laboratory testing or geomechanical analysis. However, SPSS should be interpreted as a supplementary screening tool rather than a replacement for comprehensive field-scale characterization and reservoir simulation.

The SPSS values demonstrated statistically significant differentiation across the reconstructed integrity regimes, with mean and median trends indicating progressive variation in stress–porosity behaviour among the screening classes. The relatively small variability observed within the Transitional regime suggests a comparatively constrained intermediate domain, whereas the broader distribution associated with Risk samples indicates greater heterogeneity within elevated integrity conditions. Statistical testing further confirmed significant differences through both ANOVA ($F = 90.23$, $p < 0.001$) and Kruskal–Wallis analysis ($H = 22.14$, $p < 0.001$), with large effect sizes ($\eta^2 = 0.819$; $\varepsilon^2 = 0.503$), as shown in Table 3. The agreement between parametric and non-parametric approaches suggests that SPSS captures meaningful differences in screening behaviour and may provide useful support as a simplified descriptor for preliminary caprock integrity evaluation.

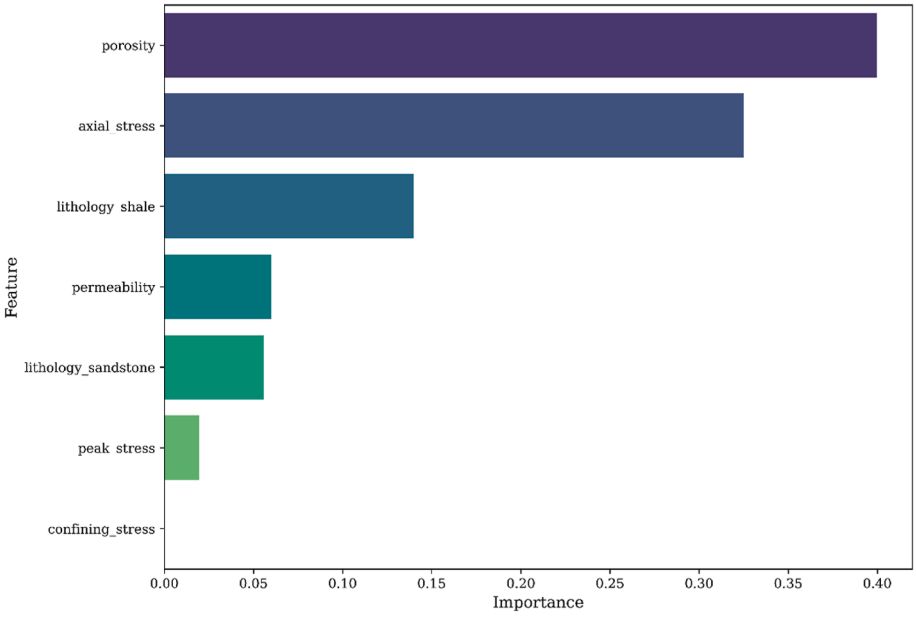


Fig. 8. Random Forest feature importance ranking for reconstructed integrity screening regimes. Consistent with SHAP interpretation, porosity and axial stress demonstrated the strongest contribution toward model prediction behaviour.

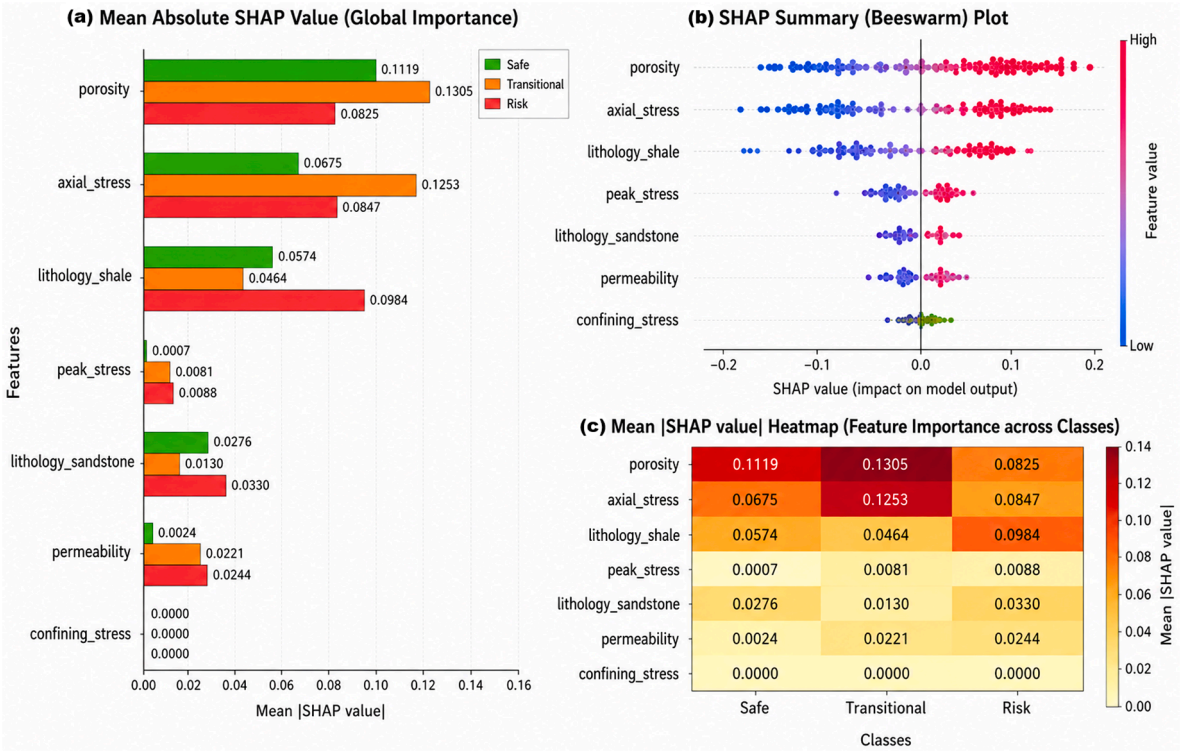


Fig. 9. SHAP-based interpretation of the Random Forest model used for caprock integrity classification. (a) Mean absolute SHAP values showing the relative importance of each predictor for the Safe, Transitional, and Risk classes. (b) SHAP summary (beeswarm) plot illustrating the distribution and direction of feature contributions across all samples, where colour represents feature magnitude and SHAP values indicate the impact on model predictions. (c) Heatmap of mean absolute SHAP values, providing a comparative visualization of feature importance across the three integrity classes.

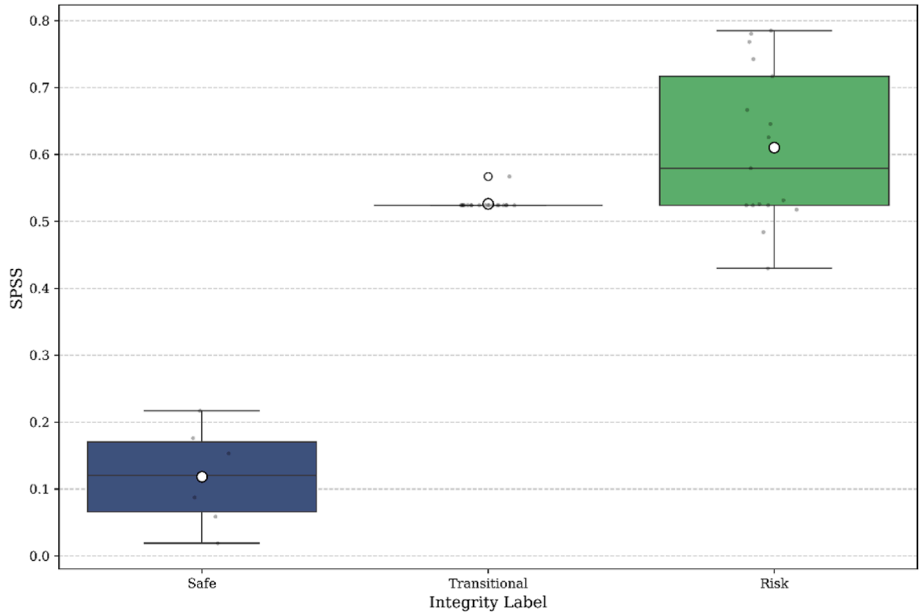


Fig. 10. Distribution of Stress–Porosity Screening Score (SPSS) across integrity regimes. Increasing SPSS values from Safe to Risk categories indicate progressive stress–porosity influence and support its capability as a preliminary screening descriptor.

4. Field implications and practical screening utility

The proposed framework demonstrates potential as a preliminary screening tool for seal integrity assessment, particularly in data-limited environments where complete geomechanical characterization may not be available. This approach may function as an early-stage decision support system capable of rapidly identifying relative integrity tendencies using commonly available parameters such as porosity, stress measurements, permeability, and lithological descriptors. This may be particularly relevant during early site evaluation stages where detailed laboratory testing, coupled simulations, or field-scale calibration remain unavailable. The reconstruction of Safe, Transitional, and Risk regimes

Table 3
Descriptive statistics and statistical validation of the Stress–Porosity Screening Score (SPSS) across reconstructed Safe, Transitional, and Risk regimes. Mean trends, distribution characteristics, and significance testing demonstrate the capability of SPSS to differentiate integrity screening behaviour among the reconstructed classes.

Integrity regime	Mean SPSS	Median SPSS	Standard deviation	ANOVA ($F = 90.23$)	Kruskal–Wallis ($H = 22.14$)	p-value	Effect size
Safe	0.1185	0.1202	0.0757	Significant	Significant	<0.001	$\eta^2 = 0.819$; $\varepsilon^2 = 0.503$
Transitional	0.5261	0.5240	0.0096	Significant	Significant	<0.001	$\eta^2 = 0.819$; $\varepsilon^2 = 0.503$
Risk	0.6101	0.5794	0.1150	Significant	Significant	<0.001	$\eta^2 = 0.819$; $\varepsilon^2 = 0.503$

Note: ANOVA ($F = 90.23$) and Kruskal–Wallis ($H = 22.14$) are overall (omnibus) statistical tests comparing all three integrity regimes simultaneously. Therefore, identical test statistics, p-values, and effect sizes are reported for each regime for completeness.

further provides a structured representation of integrity behaviour that may assist prioritization of samples or intervals for subsequent investigation.

The explainability analyses indicated that porosity and axial stress exerted stronger influence than several other variables within the present dataset, suggesting that pore structure and stress-state interactions may play an important role in defining preliminary integrity behaviour. From an operational perspective, this observation aligns with the broader understanding that storage performance and seal integrity are governed by coupled mechanical and transport processes rather than by isolated parameters. However, the identified screening behaviour should not be interpreted as evidence of universal controlling mechanisms. The relative dominance observed here remains conditioned by dataset composition, sample availability, and the selected reconstruction workflow. Additional datasets spanning broader geological settings would be necessary before generalizing these observations.

The extracted screening relationships and SPSS may also provide a pathway toward practical integration with future monitoring or risk-assessment workflows. For example, such screening outputs could be incorporated into site-ranking exercises, uncertainty reduction workflows, or integrated with reservoir and geomechanical simulations as an initial filtering layer. Nevertheless, the present framework should be viewed as a screening-oriented methodology rather than a predictive substitute for detailed subsurface characterization. The limited sample size, reconstructed labels, and absence of field-scale validation suggest that current findings are best interpreted as a proof-of-concept demonstrating the feasibility of interpretable machine learning for preliminary caprock integrity evaluation.

4.1. Limitations and future perspectives

The following limitations should be considered when interpreting the results of this study:

i. The dataset consisted of only 43 rock-mechanical samples obtained from a single CO_2 storage site. Although LOOCV and repeated cross-validation were employed to improve robustness assessment, the developed models should be regarded as preliminary screening tools and their predictive behaviour may not be directly transferable to other geological settings.
ii. The Safe, Transitional, and Risk classes were reconstructed through an unsupervised clustering workflow and ambiguity-based boundary analysis rather than being derived from observed caprock failure events, leakage records, or validated field-performance indicators. Consequently, the machine learning models developed in this study classify data-derived screening regimes rather than directly predicting caprock failure or storage security.
iii. The dataset incorporated both reservoir sandstones and caprock shales to represent the integrated reservoir–seal architecture of the Northern Lights storage system. Therefore, the reconstructed integrity regimes should be interpreted as relative screening domains within a heterogeneous rock-mechanics dataset rather than as shale-specific caprock integrity classifications.
iv. The proposed SPSS was developed as a post hoc screening descriptor based on variables identified through machine learning explainability analysis. Although SPSS demonstrated statistically significant differentiation among the reconstructed screening regimes, it has not yet been validated against independent datasets, laboratory-derived integrity criteria, or field observations.
v. The dominance of porosity and axial stress identified by Random Forest and SHAP analyses may partly reflect the characteristics and variability of the present dataset. As a result, the extracted feature rankings, decision-tree thresholds, and screening pathways should not be interpreted as universal geomechanical criteria applicable to all caprock systems.
vi. The present framework considers only static rock-mechanical and petrophysical measurements and does not explicitly account for coupled thermo-hydro-mechanical-chemical (THMC) processes, time-dependent deformation, injection-induced stress evolution, or geochemical alterations that may influence long-term caprock performance during CO_2 storage operations.
vii. Future work should focus on validating the framework using larger multi-site datasets, integrating experimentally validated caprock failure criteria, incorporating coupled THMC processes, and benchmarking the proposed workflow against conventional geomechanical modelling approaches to improve its physical relevance and practical applicability.

5. Conclusions

The following conclusions can be drawn from this study.

i. The proposed workflow provides a rapid preliminary screening approach for seal integrity assessment and may assist in early-stage CO_2 storage site evaluation where detailed geomechanical characterization is unavailable or limited.
ii. Reconstruction of Safe, Transitional, and Risk regimes allows identification of intermediate caprock conditions, which may help prioritize zones requiring further laboratory testing, monitoring, or high-fidelity numerical simulation.
iii. Random Forest demonstrated stable predictive behaviour across multiple validation strategies, suggesting its suitability as a practical screening tool for identifying integrity trends under limited and heterogeneous datasets.
iv. Explainability analysis consistently highlighted porosity and axial stress as dominant predictors within the reconstructed screening framework, indicating that stress–porosity interactions should receive greater attention during caprock characterization and data acquisition programs.
v. The SPSS offers a simplified post hoc descriptor that can support rapid comparison of caprock conditions and provide an interpretable indicator for preliminary risk prioritization.
vi. Although the framework is not intended to replace reservoir-scale geomechanical simulations, it can function as a decision-support layer to guide data collection strategies, reduce screening uncertainty, and optimize candidate selection during early CCS project planning stages.

Dataset availability and interpretation

The dataset is available in opensource repository at https://co2datashare.org/dataset/rock-mechanics-data-for-31-5-7-eos-well-core and is characterized by Griffiths et al. (2025) (Griffiths et al., 2025).

Code availability

The code utilized and fullpython script (with additional comments and tables) used in this work are available at view only Google Colab repository:https://colab.research.google.com/drive/1CSjH2Sm3A_AaVRNf2-boHMJxxf2LdUMw?usp=sharing and at Github:https://github.com/HDRASOOL/ML-based_Seal_Integrity

CRediT authorship contribution statement

Muhammad Hammad Rasool: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Writing – original draft. **A.K.M. Eahsanul Haque:** Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Numair Ahmed Siddiqui:** Investigation, Project administration, Validation, Writing – review & editing. **Hassan Soleimani:** Validation, Visualization, Writing – review & editing. **Maqsood Ahmad:** Methodology, Supervision, Writing – review & editing. **Nisar Ahmed:** Validation, Visualization, Writing – review & editing. **Muhammad Mubashir:** Validation, Writing – review & editing. **Venkateshwaran Baskaran:** Validation, Visualization, Writing – review & editing. **Md Nasirul Islam:** Validation, Writing – review & editing. **Lawal Opeyemi Olajid:** Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the Malaysia Thailand Joint Development Area (MTDJA) (Project Cost Centre No. 015QA0-044). The authors sincerely thank the Universiti Teknologi PETRONAS (UTP) Programme Office for their continuous administrative support and efficient management of the MTJDA research grant, which greatly facilitated this research.

References

Ahmed, A., Kamy, S., Delshad, M., Hildebrand, L.M.A., 2025. A comprehensive review of CO_2 subsurface storage: integrity, safety, and economic viability. Energy Geosci. 100441.

Aminaho, E.N., Hossain, M., Faisal, N.H., Sanaee, R., 2025. Application of machine learning in the determination of rock brittleness for CO_2 geosequestration. Mach. Learn. Appl. 20, 100656.

Bhagat, M., Bakariya, B., 2022. Implementation of logistic regression on diabetic dataset using train-test-split, K-Fold and stratified K-Fold approach: M. Bhagat, B. Bakariya,. Natl. Acad. Sci. Lett. 45 (5), 401–404.

de Lima, A.V.S., do Nascimento Guimarães, L.J., de Sá Beserra, L.B., Horowitz, B., Rojas Tueros, J.A., 2026. Analysis of structural integrity of geological reservoirs for CO_2 storage regarding to fault reactivation: a forecasting model based on machine learning techniques. Neural Comput. Appl. 38 (9), 344.

Griffiths, L., Thompson, N., Smith, H., Skurtveit, E., Grande, L., 2025. Rock mechanical testing of core from the Eos Northern Lights CCS confirmation well and implications for the Geomechanics of North Sea CO_2 storage. Int. J. Greenh. Gas Control 144, 104371.

Hannon Jr., M.J., Esposito, R.A., 2015. Screening considerations for caprock properties in regards to commercial-scale carbon-sequestration operations. Int. J. Greenh. Gas Control 32, 213–223.

Hawez, H.K., Bakir, H., Jamal, K., Kakakhan, M., Hussein, K., Omar, M., 2025. Numerical investigation of CO_2 injection effects on shale caprock integrity: a case study of opalinus clay. Gases 5 (3), 15.

Isah, A., Tariq, Z., Mustafa, A., Mahmoud, M., Okoroafor, E.R., 2025. A review of data-driven machine learning applications in reservoir petrophysics. Arabian J. Sci. Eng. 50 (24), 20343–20377.

Karpfinger, F., Dubost, F.-X., Samir, M., Blinov, V., Khatri, G., Finko, A., Thomas, D.P., 2025. Geomechanical evaluation for cap rock integrity. In: SPE Annual Caspian Technical Conference. SPE p. D031S019R003.

Khadka, N., 2023. How leave-one-out cross validation (LOOCV) improve's model performance. Dataaspirant-A Data Sci.Port. Beginn. 5, 542–548.

Kim, H., Vilarrasa, V., Makhnenko, R.Y., 2025. Laboratory-scale assessment of CO_2 sealing potential of heterogeneous caprock. Water Resour. Res. 61 (7) p. e2025WR040339.

Lee, S.-Y., Mohamed, F.R., Lee, K.-H., McPherson, B., Balch, R., Yoon, S., 2023. Probabilistic evaluation of geomechanical risks in CO_2 storage: an exploration of caprock integrity metrics using a multilaminate model. Energies 16 (19), 6954.

Rafiei Renani, H., Cai, M., 2022. Forty-year review of the Hoek–Brown failure criterion for jointed rock masses. Rock Mech. Rock Eng. 55 (1), 439–461.

Rasool, M.H., Ahmad, M., Ayoub, M., 2023. Selecting geological formations for CO_2 storage: a comparative rating system. Sustainability 15 (8), 6599.

Rasool, M.H., Jaffari, R., Ahmad, M., Siddiqui, N.A., Junejo, A.Z., Abbas, M.A., 2025. Novel deep learning framework for efficient pressure zone detection via analysis of pore pressure profiling. Arabian J. Sci. Eng. 50 (1), 237–248.

Ross, A., Alexander, D., 2026. A caprock integrity assessment to optimize CO_2 storage using coupled geomechanics: a case study of the Angostura field. Deep Resourc. Eng. 100252.

Sadiq, M., Sharma, R., Bisht, B.S., Chaudhary, M., 2025. Ensuring long-term caprock integrity in carbon capture, utilisation and storage: a comprehensive review of geo-mechanical and monitoring approaches. Curr. Sci. 129 (2).

Skerbisch, L., et al., 2025. Caprock integrity: a critical factor for carbon capture and storage in the Vienna Basin. Int. J. Greenh. Gas Control 146, 104434.

Suhaidi, N.N.S., Kaylyn, F.F., Ismail, N., 2025. Assessment of caprock integrity utilising rock failure prediction methodology for carbon capture and storage (CCS): a review. Adv. Mater. Res. 1184, 37–47.

Sun, Z., Salazar-Tio, R., Wu, L., Bostrøm, B., Fager, A., Crouse, B., 2023. Geomechanical assessment of a large-scale CO_2 storage and insights from uncertainty analysis. Geoenergy Sci. Eng. 224, 211596.

Tillero, E., 2024. Machine learning-based modelling for geologic CO_2 storage in deep saline aquifers. Case study of bunter sandstone in Southern North Sea. Int. J. Greenh. Gas Control 133, 104077.

Venkateshwaran, B., Haque, A.E., Ahmed, N.S., Ramkumar, M., Olutoki, J.O., 2023. A hybrid approach to caprock potential assessment and uncertainty estimation: bagging SVM and random forest. In: 2nd EAGE Workshop on Quantifying Uncertainty in Depth Imaging, vol 2023. European Association of Geoscientists & Engineers, pp. 1–7 no. 1.

Widodo, S., Brawijaya, H., Samudi, S., 2022. Stratified K-fold cross validation optimization on machine learning for prediction. Sinkron: jurnal dan penelitian teknik informatika 6 (4), 2407–2414.

Yao, P., Yu, Z., Zhang, Y., Xu, T., 2023. Application of machine learning in carbon capture and storage: an in-depth insight from the perspective of geoscience. Fuel 333, 126296.

Younessi, A., Kårstad, E., Basu, P., Larsen, M., Burachok, O., 2024. Geomechanical analysis of caprock integrity and fault stability for Greensand CO_2 storage project feasibility. In: ARMA US Rock Mechanics/Geomechanics Symposium. ARMA p. D032S038R001.