Journal of Integrated Science, Technology and Management, 2026; 2(3):29-34

Journal of Integrated Science, Technology and Management

Research Article

Explainable AI for Risk Assessment in Geological CO_2 Storage

Mohammad Hasan*

Kalam Institute of Health Sciences

ARTICLE INFO

Article history:

Received: 05 November, 2026

Revised: 20 November, 2026

Accepted: 10 December, 2026

Published: 27 December, 2026

Keywords:

Explainable artificial intelligence, geological CO_2 storage, SHAP, interpretable machine learning, risk assessment, geomechanics, carbon capture and storage

ABSTRACT

Geological CO_2 storage (GCS) is a cornerstone technology for large-scale decarbonization, but its safe and durable deployment depends on rigorous, defensible risk assessment across containment, injectivity, and geomechanical stability. Machine learning and deep learning models are increasingly used to accelerate this risk assessment, offering predictive power that often exceeds conventional statistical or deterministic methods. However, the opacity of many high-performing models—particularly deep neural networks—creates a fundamental barrier to regulatory acceptance, operator trust, and defensible decision-making in a domain where errors carry substantial environmental and financial consequences. Explainable artificial intelligence (XAI) has emerged as a critical complement to predictive modeling in GCS, providing tools that decompose model outputs into interpretable, feature-level contributions and thereby bridge the gap between black-box predictive accuracy and the transparency demanded by engineers and regulators. This review examines the theoretical foundations and applications of XAI—including SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and inherently interpretable model architectures—for risk assessment across the CO_2 storage lifecycle, drawing on recent case studies in heterogeneous depleted gas reservoir optimization, geothermal-analogue reservoir characterization, and geomechanically informed wellbore stability assessment. It further considers cross-domain parallels with explainability-driven AI frameworks developed for public health and healthcare administration, which offer transferable lessons on deploying auditable AI systems in regulated, high-consequence domains. The review concludes by identifying persistent challenges in explanation fidelity, computational cost, and the translation of feature attributions into physically meaningful engineering insight, and proposes a research agenda for advancing XAI toward routine use in commercial-scale CO_2 storage risk management.

Introduction

The scale of deployment required for geological CO_2 storage (GCS) to meaningfully contribute to global decarbonization—storing gigatonnes of CO_2 annually within depleted hydrocarbon reservoirs, deep saline aquifers, and other subsurface formations—demands risk assessment tools capable of rapidly and reliably evaluating containment integrity, injectivity performance, and geomechanical stability across highly heterogeneous geologic settings. Machine learning and deep learning models have increasingly been adopted to support this risk assessment, offering the ability to learn complex nonlinear relationships between reservoir properties, operational parameters, and risk outcomes directly from data, often at a fraction of the computational cost of full-physics simulation.

However, the predictive power of these models frequently comes at the cost of interpretability. Deep neural networks, gradient-boosted ensembles, and other high-capacity architectures function largely as "black boxes," producing accurate predictions without revealing the underlying reasoning that connects specific input features—porosity, injection rate, fault proximity, in-situ stress—to a given risk output. This opacity poses a fundamental barrier in GCS, where regulatory bodies, operators, and the public require not merely accurate

***Corresponding Author:** Mohammad Hasan

Address: Kalam Institute of Health Sciences

Email ✉**:** Mohammad.hasan@kiht.in

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

predictions but defensible, auditable justification for risk-informed decisions governing multimillion-dollar injection strategies and long-term containment commitments.

Explainable artificial intelligence (XAI) has emerged as the methodological response to this challenge, encompassing a family of techniques designed to render machine learning predictions interpretable without necessarily sacrificing predictive accuracy. These techniques range from post-hoc explanation methods—such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)—that decompose black-box model outputs into feature-level contributions, to inherently interpretable model architectures, such as decision trees and rule-based systems, that are transparent by design. This review synthesizes recent applications of XAI to risk assessment across the CO_2 storage lifecycle, integrating insights from heterogeneous reservoir optimization, geothermal-analogue reservoir characterization, and machine-learning-enhanced geomechanical risk assessment, alongside the growing body of XAI literature applied directly to CO_2 leakage detection, storage property prediction, and caprock integrity screening (Erdinc et al., 2022; Nait Amar et al., 2025; Shokrollahi et al., 2024). It further draws methodological parallels with explainability-driven AI frameworks developed for public health and healthcare administration (Nguyen, 2026a, 2026b, 2026c), which offer transferable governance lessons for deploying auditable AI systems in other regulated, high-consequence domains. The objective is to provide a comprehensive synthesis for researchers, engineers, and regulators seeking to advance XAI from a research interest toward a standard component of operational CO_2 storage risk management.

The Interpretability Imperative in Geological CO_2 Storage

Why Black-Box Models Are Insufficient for High-Consequence Decisions

Risk assessment in GCS informs decisions with substantial and often irreversible consequences: well placement, injection rate limits, monitoring network design, and long-term liability commitments. A model that predicts elevated leakage risk or fault reactivation potential without explaining which geologic or operational factors drive that prediction offers limited actionable value to an engineer who must decide how to mitigate the identified risk. Moreover, regulatory frameworks governing CCUS projects increasingly require operators to demonstrate a defensible basis for risk-informed monitoring and injection plans, a requirement that purely black-box predictive models struggle to satisfy.

The Trust Gap Between Predictive Accuracy and Operational Confidence

A persistent tension in applied machine learning for subsurface engineering is that the most predictively accurate models—deep convolutional and recurrent architectures, gradient-boosted ensembles—are often the least interpretable, while inherently interpretable models—linear regression, shallow decision trees—sacrifice predictive performance for transparency. XAI techniques aim to resolve this tension by extracting interpretable explanations from high-capacity models post hoc, allowing practitioners to retain predictive accuracy while gaining the feature-level insight needed to build operational trust and validate that a model's reasoning is consistent with established geologic and engineering understanding.

Regulatory and Stakeholder Requirements for Auditability

Beyond engineering utility, explainability serves a governance function: it allows regulators, investors, and communities near storage sites to audit the basis for risk assessments that inform permitting and monitoring decisions. As machine learning becomes further embedded in CO_2 storage risk workflows, the capacity to trace a given risk alert or capacity estimate back to specific, physically interpretable drivers becomes essential to sustaining the social and regulatory license required for continued CCUS deployment.

Foundations of Explainable AI Techniques

SHapley Additive exPlanations (SHAP)

SHAP, grounded in cooperative game theory's Shapley value framework, quantifies each input feature's contribution to a model's output by considering its marginal contribution across all possible feature coalitions, providing both global feature importance rankings and instance-specific explanations for individual predictions (Lundberg & Lee, 2017). SHAP has become the dominant XAI technique in subsurface energy applications due to its strong theoretical guarantees of consistency and local accuracy. Shokrollahi et al. (2024) applied SHAP analysis to a random forest model predicting CO_2 solubility in brine—a key parameter governing solubility trapping efficiency in saline aquifer storage—demonstrating that while pressure and temperature were the dominant predictive factors, salinity and specific ionic species, notably chloride, contributed materially to accurate predictions, insights that would remain hidden within an unexplained black-box solubility model.

Local Interpretable Model-Agnostic Explanations (LIME)

LIME generates local, interpretable approximations of a complex model's behavior in the neighborhood of a specific prediction, fitting a simple, interpretable surrogate model to explain individual predictions even when the underlying model is highly nonlinear. Nait Amar et al. (2025) combined LIME and SHAP interpretability techniques in developing explainable machine learning models for predicting CO_2-brine interfacial tension—a parameter directly governing capillary trapping efficiency and structural containment capacity—confirming that the extracted explanations aligned with established physicochemical trends and thereby enhancing confidence in the model's applicability to novel sequestration depth and salinity conditions.

Class Activation Mapping and Saliency-Based Explanations for Spatial Data

For spatially structured data such as time-lapse seismic images, class activation mapping (CAM) and related saliency-based techniques localize the image regions most influential to a convolutional neural network's classification decision, providing a visual explanation of which spatial features drove a given leakage detection outcome. Erdinc et al. (2022) developed an explainable CO_2 leakage detection framework that combined binary classification of time-lapse seismic monitoring images with class activation mapping to localize the specific leakage region driving a positive detection, directly addressing the operational need to translate a leakage alert into an actionable, spatially localized response rather than an unexplained binary flag.

Inherently Interpretable Architectures

An alternative to post-hoc explanation of black-box models is the use of inherently interpretable architectures—decision trees, rule-based systems, and sparse linear models—that are transparent by construction. Recent work on interpretable machine learning for preliminary caprock seal integrity screening has demonstrated that decision-tree-based frameworks, informed by feature importance and SHAP analysis, can extract data-emergent screening thresholds that translate learned relationships into transparent, engineering-usable decision pathways, offering a complementary approach to purely post-hoc explanation of more complex model architectures.

Applications to Reservoir Characterization and Risk Assessment

Explainable Petrophysical Property Prediction

Machine learning-enhanced petrophysical workflows applied to the Three Forks geothermal reservoir in the Williston Basin illustrate how well log, core, and seismic attribute data can be integrated to predict reservoir quality at high spatial resolution. Embedding XAI techniques within such workflows—for example, applying SHAP analysis to identify which log responses or seismic attributes most strongly drive a given porosity or permeability prediction—would allow geoscientists to validate that model predictions are grounded in established petrophysical relationships rather than spurious data correlations, an important safeguard given the consequential role these property predictions play in downstream CO_2 storage capacity and injectivity estimates.

Explainable Geomechanical Risk Assessment

Machine-learning-enhanced geomechanical characterization and wellbore stability assessment, as demonstrated in the offshore South Tano Field, integrates stress field estimation, rock strength prediction, and drilling parameter analysis into a proactive risk-management workflow. Applying explainability techniques to this class of geomechanical model is particularly valuable given the safety-critical nature of wellbore stability predictions: a SHAP-based decomposition of a stability risk score, for instance, could reveal whether a predicted instability is driven primarily by in-situ stress anisotropy, rock strength heterogeneity, or drilling parameter deviation, directly informing the choice of mitigation strategy rather than merely flagging elevated risk.

Explainable Caprock and Seal Integrity Screening

Interpretable machine learning frameworks for preliminary seal integrity screening illustrate a workflow in which unsupervised clustering first reconstructs safe, transitional, and at-risk integrity regimes from limited rock mechanics data, after which supervised classifiers—evaluated for both predictive performance and explainability—are used to identify the dominant geomechanical controls governing caprock integrity. SHAP-based analyses of such models have consistently identified porosity and failure-related effective axial stress as dominant controls on seal integrity classification, providing a transparent, data-adaptive basis for prioritizing detailed geomechanical investigation at specific storage sites, including heterogeneous settings such as the Evetts site in the depleted-gas Permian Basin.

Explainable Storage Capacity and Trapping Mechanism Prediction

Beyond containment risk, XAI techniques are increasingly applied to explain predictions of storage capacity and trapping efficiency—outcomes governed by solubility trapping, capillary trapping, and structural trapping mechanisms whose relative contributions vary substantially across geologic settings. By decomposing capacity or trapping-efficiency predictions into feature-

level contributions from pressure, temperature, salinity, and interfacial properties (Nait Amar et al., 2025; Shokrollahi et al., 2024), explainable models allow operators to identify which site-specific parameters most strongly govern storage performance, informing site selection and injection design decisions with a level of physical insight that purely predictive models cannot provide.

Explainability Across the Risk Assessment Workflow

Pre-Injection Site Screening

At the site screening stage, explainable models can support transparent ranking of candidate storage sites by decomposing composite risk or suitability scores into their constituent geologic and geomechanical drivers, enabling operators to communicate the basis for site selection decisions to regulators and stakeholders in physically meaningful terms rather than as an opaque composite score.

Injection Design and Real-Time Monitoring

During active injection, explainable anomaly detection models—such as the class-activation-mapping-based leakage localization approach demonstrated by Erdinc et al. (2022)—provide operators with spatially and physically grounded explanations of detected anomalies, supporting faster and more targeted operational response than an unexplained binary leakage alert would allow.

Post-Injection Containment Verification

In the post-injection monitoring phase, explainable models supporting long-term containment verification must be able to justify their risk assessments over timescales spanning decades, during which the underlying geologic and operational conditions may drift outside the range represented in original training data; explainability techniques that flag when a given risk prediction is poorly supported by the available feature evidence—rather than only explaining well-supported predictions—represent an important and still-developing capability for this stage of the CO_2 storage lifecycle.

Cross-1Domain Lessons: Explainability-Driven AI Governance in Public Health and Healthcare Administration

While the technical foundations of XAI for GCS risk assessment are grounded in subsurface engineering, valuable governance and deployment lessons can be drawn from explainability-driven AI frameworks developed in other high-stakes, regulated domains. National-scale AI frameworks developed for chronic disease prevention and early intervention illustrate the operational necessity of explainable, clinically grounded risk stratification: a model that flags a patient as high-risk without indicating which clinical or behavioral factors drove that assessment offers limited actionable value to a care team and risks eroding clinician trust (Nguyen, 2026a). This principle—that risk predictions must be accompanied by interpretable, domain-grounded justification to support actionable intervention—maps directly onto CO_2 storage risk assessment, where a geomechanical or containment risk flag similarly requires feature-level explanation to translate into an effective mitigation response.

Frameworks for transforming prior authorization through artificial intelligence further illustrate how explainability supports institutional auditability: automated decisions affecting patient access must be traceable to specific, policy-consistent justifications to satisfy regulatory and administrative review (Nguyen, 2026b). This parallels the auditability requirements surrounding GCS monitoring, verification, and accounting (MVA) reporting, where risk assessments generated by machine learning models must similarly be traceable to specific, physically justified drivers to satisfy regulatory review—an argument for treating explainability not as an optional model enhancement but as a prerequisite for regulatory-grade deployment of AI-driven risk assessment in CCUS.

Economic impact analyses of AI adoption in healthcare demonstrate that explainability-driven AI systems—by supporting more targeted, trusted intervention—can generate measurable cost savings and productivity gains relative to opaque decision-support tools that clinicians and administrators are reluctant to fully trust (Nguyen, 2026c). An analogous economic case can be constructed for XAI-based CO_2 storage risk assessment: models whose outputs are explainable and therefore more readily trusted and acted upon by operators and regulators are likely to reduce costs associated with excessive precautionary monitoring, delayed permitting, or reactive rather than proactive risk mitigation, an economic argument that merits explicit quantification as XAI systems mature toward routine CCUS deployment.

Collectively, these cross-domain parallels reinforce a broader principle: explainability is not a peripheral technical feature but a central requirement for trustworthy AI deployment in domains where erroneous or opaque predictions carry substantial financial, safety, or public-health consequences. The maturation of XAI for GCS risk assessment can accordingly draw on governance, validation, and stakeholder-communication frameworks already developed in these adjacent, more mature AI application domains.

Challenges and Limitations

Despite substantial progress, several challenges continue to constrain the operational deployment of XAI for GCS risk assessment.

Explanation fidelity and stability. Post-hoc explanation methods such as SHAP and LIME approximate a complex model's behavior, and their explanations can be sensitive to the choice of background data, sampling strategy, or perturbation scheme, raising questions about the stability and fidelity of the resulting feature attributions, particularly for highly nonlinear deep learning architectures.

Computational cost. Exact Shapley value computation scales exponentially with the number of input features, and while efficient approximation algorithms exist, applying SHAP-based explanation to high-dimensional spatial or spatiotemporal data—such as full seismic image volumes or dense reservoir property grids—remains computationally demanding relative to the underlying predictive model itself.

Translating feature attributions into physical insight. A statistically significant feature attribution does not automatically constitute a physically meaningful engineering explanation; bridging the gap between a SHAP value and an actionable geologic or geomechanical interpretation requires domain expertise that is not automatically encoded within the explanation method itself.

Explaining well-calibrated uncertainty, not just point predictions. Most current XAI techniques explain a model's point prediction rather than its associated predictive uncertainty, limiting their utility in risk assessment contexts where understanding the confidence and evidentiary support behind a risk estimate is as important as understanding its point value.

Standardization and regulatory acceptance. No standardized protocol currently exists for validating or benchmarking XAI explanations in the specific context of subsurface risk assessment, complicating efforts to establish explainability as a formal, auditable component of regulatory review for CCUS projects.

Future Directions

Future research should prioritize the development of explanation methods specifically tailored to the spatiotemporal and multiphysics character of GCS risk assessment, extending beyond tabular SHAP and LIME applications (Nait Amar et al., 2025; Shokrollahi et al., 2024) toward explanation frameworks natively suited to convolutional and neural operator architectures used for plume migration and geomechanical prediction. Building on the localized, spatially grounded explanation approach demonstrated for seismic leakage detection (Erdinc et al., 2022), extending class-activation and saliency-based explanation techniques to coupled flow-geomechanics and reactive transport surrogate models represents a promising avenue for improving the interpretability of the full range of risk assessment models relevant to heterogeneous storage sites such as the Evetts site.

Integration of explainability techniques directly into high-resolution petrophysical and geomechanical characterization workflows—such as those demonstrated for the Three Forks geothermal reservoir and the South Tano offshore field—represents a further avenue for improving the transparency and defensibility of site-specific risk assessment. Finally, adopting governance and validation frameworks informed by explainability-driven AI deployment in public health and healthcare administration (Nguyen, 2026a, 2026b, 2026c) may help establish the standardized auditability protocols required for XAI systems to transition from research demonstration to routine, regulator-accepted use in commercial-scale CO_2 storage risk management.

Conclusion

Explainable artificial intelligence represents an essential complement to predictive machine learning for geological CO_2 storage risk assessment, resolving the tension between the predictive power of high-capacity models and the transparency required for defensible, regulator- and operator-trusted decision-making in a high-consequence subsurface domain. Applications spanning SHAP-based petrophysical and thermodynamic property explanation, LIME-based interfacial property interpretation, class-activation-based leakage localization, and interpretable caprock integrity screening collectively demonstrate that embedding explainability into machine learning risk assessment workflows substantially improves both engineering actionability and stakeholder trust relative to unexplained black-box predictions alone. Cross-domain lessons from explainability-driven AI governance in public health and healthcare administration further suggest a pathway toward the standardized validation and auditability frameworks necessary for XAI systems to achieve regulatory acceptance and routine operational adoption. Realizing the full potential of explainable AI in this domain will require sustained investment in explanation fidelity, computational efficiency, and the translation of feature attributions into physically grounded engineering insight—investments essential to ensuring that geological CO_2 storage can be assessed, monitored, and managed safely, transparently, and at the scale required to meet global climate objectives.

References

1. Erdinc, H. T., Gahlot, A. P., Yin, Z., Louboutin, M., & Herrmann, F. J. (2022). De-risking carbon capture and sequestration with explainable CO2 leakage detection in time-lapse seismic monitoring images. In *Proceedings of the AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*. Association for

the Advancement of Artificial Intelligence. https://arxiv.org/abs/2212.08596
2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
3. Nait Amar, M., Youcefi, M. R., Alqahtani, F. M., Djema, H., & Ghasemi, M. (2025). Rigorous explainable artificial intelligence models for predicting CO2–brine interfacial tension: Implications for CO2 sequestration in saline aquifers. *Energy & Fuels, 39*(29), 14237–14253. https://doi.org/10.1021/acs.energyfuels.5c01489
4. Nguyen, T. T. (2026a). AI-powered precision public health: A national framework for targeting chronic disease prevention and early intervention. *International Journal of Emerging Trends in Computer Science and Information Technology, 7*(3), 10–20.
5. Nguyen, T. T. (2026b). Transforming prior authorization through artificial intelligence: A national framework for reducing administrative burden and improving patient access. *International Journal of AI, BigData, Computational and Management Studies, 7*(3), 17–27.
6. Nguyen, T. T. (2026c). The economic impact of AI adoption in healthcare: Estimating national cost savings, productivity gains, and long-term health outcomes. *International Journal of Artificial Intelligence, Data Science, and Machine Learning, 7*(3), 1–11.
7. Shokrollahi, A., Tatar, A., & Zeinijahromi, A. (2024). Advancing CO2 solubility prediction in brine solutions with explainable artificial intelligence for sustainable subsurface storage. *Sustainability, 16*(17), 7273. https://doi.org/10.3390/su16177273

HOW TO CITE THIS ARTICLE: Hasan, M. (2025). Explainable AI for Risk Assessment in Geological CO_2 Storage. *Journal of Integrated Science, Technology and Management*, 2(3), 29-34.