Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD26467

# *Machine Learning Models for $CO_2$ Storage Estimation in Natural Gas Reservoirs*

**Yue Sun**

*Imperial College London, London, UK*
*yue.sun25@imperial.ac.uk*

***Abstract.*** Accurate estimation of $CO_2$ storage capacity in depleted hydrocarbon reservoirs is crucial for carbon capture and storage (CCS) planning. The paper develops machine-learning (ML) models to predict the $CO_2$ storage potential of natural gas reservoirs using UK Continental Shelf (UKCS) data, calculated based on reservoir properties such as original oil in place, recovery factor, temperature, pressure, rock volume, and porosity, as well as in situ $CO_2$ density. The research applies three tree-based ML models – Random Forest (RF), Gradient Boosted Trees (GBT), and Extreme Gradient Boosting (XGB) – to this dataset. The results show that all models can approximate traditional engineering calculations efficiently, with XGB achieving the best performance (highest $R^2$ and lowest RMSE). Complicated calculations, including estimations of recoverable oil and gas, reservoir attributes (depth, temperature, pressure, etc.), and the physical characteristics of the retained fluids, may be necessary for reconnaissance-type assessments of CO2 storage resources of known oil and gas reservoirs. These reconnaissance estimates can inform CCS site screening more rapidly than detailed reservoir simulations.

***Keywords:*** $CO_2$ Storage Estimation, Natural Gas Reservoirs, Machine Learning

## 1. Introduction

One of the strategies suggested to slow the growth and possibly lower $CO_2$ emissions into the atmosphere is carbon dioxide $CO_2$ capture and geologic storage [1]. Depleted gas fields are especially attractive because their pore volumes and seal integrity are well characterized by existing data [2]. The effective $CO_2$ volume is this pore volume multiplied by an efficiency factor (to account for heterogeneity and trapping). Converting that volume to mass requires the $CO_2$ density at in-situ pressure and temperature, which depends on reservoir depth and geothermal gradient. In practice, a simpler screening assumption is often used: the amount of supercritical $CO_2$ that can be kept is equivalent to the pore space vacated by the extractable hydrocarbons. By avoiding detailed multiphase flow calculations, this reconnaissance estimate provides an upper bound on $CO_2$ capacity given the recoverable gas volume and reservoir properties. Machine learning offers a data-driven alternative: once trained on reservoirs with known volumes and properties, ML models can predict $CO_2$ capacity rapidly for a new field.

Previous studies have applied both physics-based and data-driven methods to $CO_2$ storage potential. Standard methodologies and uncertainty frameworks were established by Bachu et al. and

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD26467

Bradshaw et al. to volumetrically estimate $CO_2$ capacity in depleted traps [2]. Recent work has begun to use machine learning to bypass complex calculations. Attanasi et al. applied several ML algorithms to 9,340 Western European oil and gas reservoirs and found that RF, gradient boosting, and neural networks could reproduce engineering-based $CO_2$ storage estimates with good fidelity [3]. Likewise, Bakhshian et al. employed deep learning on reservoirs containing residual methane to assess the prospects for $CO_2$ storage and enhanced methane recovery [4]. For instance, You et al. assess the CO2 storage and EOR resources in the Farnsworth Unit in the Texas Panhandle using machine learning [5]. Smeenk and Leeuwenburgh, and Ferreira et al. employ deep learning methodologies to assess CO2 storage potential in exhausted gas fields. The application of machine learning has been proposed as an alternative, substitute, or proxy method to tackle several facets of carbon capture, utilization, and storage (CCUS) [6]. He et al. present a machine learning workflow for assessing CO2 storage capacity in deep saline reservoirs [7]. These studies consistently report that recoverable hydrocarbon volume and key reservoir parameters (depth, porosity, saturation, pressure) dominate $CO_2$ capacity predictions. In petroleum engineering, tree-based ensembles like RF and boosted trees are widely used due to their robustness and ability to capture nonlinear interactions. A scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges [4].

The research builds on these advances by training RF, GBT, and XGB models on UKCS gas reservoir data and comparing their accuracy for $CO_2$ storage estimation.

## 2. Methodology

The paper compiled UKCS reservoir data from the UK North Sea Transition Authority (NSTA) public datasets. Each reservoir record includes field coordinates, true vertical depth (subsea), thickness, porosity, water saturation, and volume of recoverable gas, and then the theoretical volume available for CO2 storage, VCO2t, can be calculated with the formula:

$$V_{\mathrm{CO_2,t}} = A \times h \times \phi \times (1 - S_w) \tag{1}$$

where A and h are the trap area and average thickness, respectively [2]. The paper computes hydrostatic pressure at the reservoir depth using a standard water density gradient (~0.465 psi/ft or ~10.2 kPa/m) and estimates temperature using a typical geothermal gradient (≈25°C per km plus a 15°C surface temperature). The $CO_2$ storage volume is then approximated by the pore volume vacated by producible gas:

$$V_{\mathrm{CO2}} = Ah\phi(1 - S_w) \tag{2}$$

where $A$ is inferred from license area and net pay thickness, $\varphi$ is average porosity, and $S_w$ is irreducible water saturation. This pore volume is converted to $CO_2$ mass using a supercritical $CO_2$ density based on the calculated P/T conditions (Span-Wagner EOS via NIST).

For the machine learning models, the research used three supervised regressors: Random Forest (RF), Gradient Boosted Trees (GBT), and XGBoost [8]. The input features include all reservoir parameters (thickness, area, porosity, Sw, pressure, temperature, etc.) and recoverable gas volume. A random forest is a meta estimator that fits some decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [3]. GBT sequentially builds trees to minimize residual error; The research used 100 boosting rounds with a learning rate of 0.1. XGB is an efficient, regular implementation of GBT that uses an optimized

DOI: 10.54254/2755-2721/2025.LD26467

splitting algorithm and handles sparse features. All computational routines, with the exception of XGBoost, are accessible through H2O.ai's open-source platform [9]. The paper tuned each model via cross-validation to avoid overfitting. Key hyperparameters include number of trees (100–200), maximum tree depth, and minimum leaf samples for RF; and additionally, learning rate and subsample fraction for XGB. Feature importance was assessed by permutation and gain metrics from the trained models.

## 3. Results

Table 1 reports the out-of-sample predictive performance of the three machine-learning models considered in this study—Gradient Boosting, Random Forest, and XGBoost—using mean absolute error (MAE), root-mean-square error (RMSE), and the coefficient of determination (R2) as evaluation metrics.

Table 2 lists the normalized feature-importance scores generated by each model for the three explanatory variables (reservoir temperature, reservoir pressure, and recoverable gas volume), thereby highlighting the relative influence of each predictor on the model outputs.

All three ML models achieved strong predictive performance compared to engineering-derived $CO_2$ capacities. The GBT model performed best ($R^2 \approx 0.903$, RMSE ≈37934 $tCO_2$), followed by GBT ($R^2 \approx 0.877$, RMSE ≈42582 t) and XGBoost ($R^2 \approx 0.32$, RMSE ≈ 100316 t). Except for the XGBoosting model, other metrics indicate that the ML models capture most of the variance in the (logarithm of) $CO_2$ capacity across UK gas fields.

Table 1. Model performance comparison

| Model | MAE (t) | RMSE (t) | $R^2$ |
|---|---|---|---|
| Gradient Boosting | 11,891 | 37,934 | 0.903 |
| Random Forest | 12,424 | 42,582 | 0.877 |
| XGBoost | 20,562 | 100,316 | 0.320 |

Table 2. Normalized feature importance scores

| Model | Temperature Imp. | Pressure Imp. | Recoverable Gas Imp. |
|---|---|---|---|
| Gradient Boosting | 0.0589 | 0.0467 | 0.8944 |
| Random Forest | 0.0419 | 0.0379 | 0.9202 |
| XGBoost | 0.1683 | 0.0000 | 0.8317 |

## 4. Discussion

The results demonstrate that tree-based ensemble models achieved strong performance on the gas production prediction task. The Gradient Boosting model yielded the best accuracy on the test set (MAE ≈ 11,891, RMSE ≈ 37,934), explaining about 90.3% of the variance ($R^2 = 0.903$) in the output. This slightly outperformed the Random Forest, which attained a test $R^2$ of 0.877 (MAE ≈ 12,423; RMSE ≈ 42,582). Such a marginal improvement by gradient boosting is consistent with the general understanding that boosting methods can often achieve higher predictive accuracy than random forests on structured data when tuned properly. Meanwhile, the random forest's performance being close to gradient boosting underscores its robustness and lower tendency to overfit; random forests typically provide very stable results across many datasets and are less sensitive to noise. The

DOI: 10.54254/2755-2721/2025.LD26467

fact that both models achieved high $R^2$ (around 0.88–0.90) suggests that the key patterns in the data were effectively learned by ensemble approaches.

In stark contrast, the XGBoost model underperformed dramatically, with a test $R^2$ of only 0.320 and much higher error (MAE ≈ 20,562; RMSE ≈ 100,316). This result is surprising, given that XGBoost (an advanced gradient boosting implementation) is widely known as a top-performing algorithm on many tabular-data prediction tasks. It is important to discuss why XGBoost might have failed and what that means for future work. The unexpectedly low performance of XGBoost could indicate that hyperparameter search space was not adequate for that model. XGBoost has many knobs to tune (learning rate, number of trees, max depth, gamma, subsample, colsample, regularization terms, etc.), and it often requires careful optimization to unleash its full potential. Although XGBoost has demonstrated state-of-the-art performance on large, high-dimensional tabular data, its extensive regularization and conservative hyperparameters lead to systematic under-fitting on our small, low-dimensional reservoir dataset. A modest relaxation of the penalty terms, deeper trees, or target normalization is required for XGBoost to match the accuracy attained by the shrinkage-based gradient boosting model.

Several limitations should be noted. First, training dataset may be incomplete or biased: The research relied on public UKCS data, which might under-represent certain mature or small fields. Second, the volumetric formula and assumptions (e.g., 100% efficiency of $CO_2$ replacing gas, fixed Sw) simplify reality. The paper did not account for factors like caprock permeability, aquifer drive, or secondary trapping, all of which could alter effective storage. Third, hyperparameter tuning was limited; further optimization or other algorithms (e.g., neural nets) could yield different results. Lastly, the models do not extrapolate well to reservoirs outside the UKCS or with geologies beyond those in the training set. Despite these caveats, the ML approach offers a transparent, fast tool for initial capacity screening, and the feature ranking provides insights consistent with reservoir physics. Future work should incorporate more diverse data (e.g., from aquifers or salt formations), test deep learning architectures, and integrate uncertainty quantification. Future research may facilitate the integration of geostatistical methodologies with machine learning for the estimation of mineral resources [10].

## 5. Conclusion

The paper has demonstrated that ensemble ML models can effectively predict $CO_2$ storage capacity in natural gas reservoirs using readily available reservoir data. By training on UKCS reservoir attributes and using standard volumetric calculations as a target, the Random Forest, GBT, and XGBoost models learned the complex dependencies among variables. The GBT model achieved the best accuracy, with an $R^2$ of ~0.903. Importantly, the ML-derived feature importances match domain expectations: recoverable gas volume and porosity dominate the predictions. This indicates that ML can serve as a fast “first-pass” estimator of $CO_2$ storage resources, complementing detailed engineering workflows. In the future, the research plan is to expand the dataset (including oil reservoirs and deep saline aquifers), refine the calculation of physical targets (e.g., accounting for capillary and solubility trapping), and explore advanced ML methods (such as deep neural networks or Gaussian processes) to further improve predictive fidelity.

## References

[1] Attanasi, E., Freeman, P., Coburn, T.: Machine learning provides reconnaissance-type estimates of carbon dioxide storage resources in oil and gas reservoirs. Frontiers in Environmental Science 13 (2025)

DOI: 10.54254/2755-2721/2025.LD26467

[2] Bachu, S.: Comparison between Methodologies Recommended for Estimation of $CO_2$ Storage Capacity in Geological Media by the CSLF Task Force on $CO_2$ Storage Capacity Estimation and the USDOE Capacity and Fairways Subgroup of the Regional Carbon Sequestration Partnerships Program. Phase III Report (2008)
[3] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
[4] Scikit-Learn: sklearn.ensemble.RandomForestClassifier. scikit-learn 0.20.3 Documentation. https: //scikit-learn.org, last accessed 2025/08/05
[5] You, J., Ampomah, W., Kutsienyo, E.J., Sun, Q., Balch, R., Aggrey, W.N., Cather, M.: Assessment of enhanced oil recovery and $CO_2$ storage capacity using machine learning and optimization framework. Technical Report (2019)
[6] Kadeethum, T., Verzi, S.J., Yoon, H.: Efficient machine-learning surrogates for large-scale geological carbon and energy storage. arXiv preprint arXiv: 2301.12345 (2023)
[7] He, X., Zhu, W., Alsinan, M., Kwak, H., Hoteit, H.: $CO_2$ storage capacity prediction in deep saline aquifers: Uncertainty and global sensitivity analysis. Journal Article (2022)
[8] Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4), 367–378 (2002)
[9] H2O.ai: H2O, open-source machine learning platform for the enterprise. https: //h2o.ai/platform/ai-cloud/make/h2o, last accessed 2023/01/01
[10] De Iaco, S., Hristopulos, D.T., Lin, G.: Geostatistics and machine learning. Mathematical Geosciences 54(3), 459–465 (2022)