Machine Learning and Deep Learning for Mineralogy Interpretation and CO_2 Saturation Estimation in Geological Carbon Storage: A Case Study in the Illinois Basin

Hongsheng Wang[1, 2], Sherilyn Williams-Stroud[3], Dustin Crandall[4], and Cheng Chen[5 *]

[1] Department of Mining and Mineral Engineering, Virginia Tech, Blacksburg, VA 24061, USA

[2] Now at the Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX 78758, USA

[3] Illinois State Geological Survey, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA

[4] National Energy Technology Laboratory, Morgantown, WV 26507, USA

[5] Department of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

* corresponding author: cchen6@stevens.edu

Abstract: Carbon capture and storage (CCS) is a promising approach to simultaneously maintaining energy security and reducing carbon dioxide (CO_2) emissions under the current energy portfolio that is dominated by fossil fuel energy. Pre-injection formation characterization and post-injection CO_2 monitoring are two critical tasks to guarantee

storage efficiency in CCS. The CCS projects in the Illinois Basin, the first large-scale CO_2 injection into saline aquifers in the United States, employed conventional and the latest pulsed neuron tools, the pulsed neuron eXtreme (PNX), for mineralogy interpretation and CO_2 saturation estimation, which provide valuable references for future CCS projects. Because of the inherent fuzziness of petrophysical measurements and complex subsurface heterogeneity, interpreting well-logging data is time-consuming, and its accuracy can be user-biased. In recent years, data-driven methods have been widely used to capture the non-linear patterns between input features and interpretation results. This work applied and evaluated four commonly used machine learning (ML) models, including ridge regression (RR), random forest (RF), gradient boosting regression (GBR), support vector regression (SVR), and one deep learning (DL) model, the artificial neural network (ANN). We optimized the hyperparameters of the four ML models and the DL model using the simulated annealing algorithm and the grid search strategy, respectively. The input features of the mineralogy interpretation models were eleven conventional well-logging curves, and the label data (i.e., ground truth) were the porosity and volumetric fractions of six minerals, including quartz, feldspar, dolomite, calcite, clay, and iron minerals. The results demonstrated that the GBR and RF models were superior in predicting volumetric fractions of minerals and porosity; label data with low coefficient of variation (CV) values tended to yield better performance. For CO_2 saturation estimation, the RF is the best-performing model, followed by SVR, ANN, GBR, and RR. Furthermore, we conducted feature importance ranking using the permutation importance algorithm and found that the formation sigma

and well pressure were the most important features. The study of CCS projects in the Illinois Basin bridges the gap between the limited knowledge and understanding of geological carbon storage and the increasing demand for reliable, cost-effective, and sustainable energy solutions.

Keywords: Geological carbon storage, supervised machine learning, pulsed neutron eXtreme tool, conventional well-logging tool, Illinois Basin

1. Introduction

Over the past century, the consumptions of fossil fuel energy, such as coal, oil, and natural gas, have led to a significant increase in annual carbon dioxide (CO_2) emissions in the atmosphere. The global energy-related CO_2 emissions reached a new peak of 36.8 gigatonnes in 2022, according to the International Energy Agency (IEA, 2022; IEA, 2023). CO_2 is one of the major greenhouse gases (GHGs). The scientific consensus is that the increasing GHG concentrations in the atmosphere have led to worldwide global warming with climate changes resulting widespread public concerns about environmental issues (Lynas et al., 2021). Despite the availability of renewable energy sources such as solar, wind, and tidal power, it is not currently feasible to replace the dominance of fossil fuels in our energy systems, given the existing technological levels and industrial infrastructure (York & Bell, 2019; Holechek et al., 2022; Zastempowski, 2023). Therefore, to maintain and further reduce atmospheric CO_2 concentrations, it is urgent to adopt carbon capture and storage (CCS) technologies, which involve the

63 capture of CO_2 at the source, followed by dehydration, compression, transportation, and
64 sequestration (Boot-Handford et al., 2014). Carbon sequestration refers to the long-term
65 storage of CO_2 by physical or chemical processes (Guo et al., 2020; Ershadnia et al.,
66 2021), including geological storage (Juanes et al., 2006; Bachu, 2008), ocean storage
67 (Renforth & Henderson, 2017), and mineralization (Gaus, 2010). Among them,
68 geological storage is the most feasible approach because of the low cost, minimum
69 environmental influence, and potential economic benefits, such as enhanced oil and gas
70 production (Leung et al., 2014; Markewitz et al., 2012; Jia et al., 2019).

71 Geological storage is usually implemented in depleted oil and gas reservoirs, saline
72 aquifers, and unminable coal seams, which have adequate storage capacities and
73 impermeable sealing caprocks (Pires et al., 2011). Pre-injection formation
74 characterization and post-injection CO_2 plume monitoring are two important and
75 challenging tasks in CCS projects (Bauer et al., 2016; Dance et al., 2019) to guarantee
76 storage security, which can be augmented with petrophysical technologies, including
77 direct and indirect approaches. Laboratory core analyses such as X-ray diffraction
78 (XRD) and core-flooding experiments provide direct observation and precise property
79 characterization of formation samples (Clarkson et al., 2019; Li et al., 2020), which are
80 essential for validating petrophysical-measured properties. However, despite their
81 importance, these methods can be costly and often yield a limited number of data points,
82 thus not providing a broad enough picture of the formation (Lindberg et al., 2015). In
83 contrast, well-logging technologies provide comprehensive and spatially continuous
84 elaboration of the formation parameters along the drilling trajectory. The logging

85 measurements related to the bulk properties, such as natural radioactivity, electrical
86 conductivity, acoustic conductivity, are influenced by the rock matrix, fluids in the pore
87 space, and their interactions (Corina et al., 2018). As a result, the logging data can be
88 used to infer the static formation properties, such as the porosity, permeability,
89 mineralogy, and fluid saturation. Time-lapse logging data provides insights into the
90 dynamic subsurface fluid flow behaviors. Furthermore, the log suites, which combine
91 different logging tools responding to the same properties, can make the interpretation
92 results more reliable. Conventional well-logging tools, including resistivity, gamma ray,
93 spontaneous potential, acoustic, have been widely used in open-hole wells (Zhang, 2011;
94 Lai et al., 2018). In addition, the pulsed neutron tools expand application scenarios,
95 thereby enabling fluid saturation estimation in cased wells (Rose et al., 2015). The
96 recent advent of the pulsed neutron eXtreme (PNX) tools further improves saturation
97 measurements in the presence of gas in the pore space by employing additional
98 properties that are highly sensitive to gas volume variations. Well-logging technologies
99 play significant roles in various applications, such as lithology identification (Ren et al.,
100 2019), hydrocarbon and hydrological exploration (Kamali et al., 2004; Tan et al., 2015),
101 ore deposit location (Houseman et al., 1989; Silversides et al., 2015), reservoir
102 evaluation (Yang & Aplin, 2010), and CO_2 monitoring (Baumann et al., 2014).

103 Well-logging data analysis and interpretation involve iterative inversions to unravel
104 the non-linear relationship between the logging measurements and subsurface in-situ
105 rock-fluid properties (Wang & Carr, 2012; Kim et al., 2020). Because of the intrinsic
106 fuzzy nature of the petrophysical response and the complex subsurface heterogeneity,

the inversion process is usually time consuming and the accuracy is affected heavily by prior knowledge of the target formations (Szabó et al., 2020). For instance, the constants in the Archie's law (Cai et al., 2017), an empirical equation for the bulk electrical resistivity analysis, are pre-defined based on the formation lithology and vary with multiple factors, such as rock properties, fluid nature, and user interpretation. Data-driven models, such as machine learning (ML) and deep learning (DL) models, are capable of capturing the complicated and indistinct patterns in large multivariate datasets and making predictions for unseen datasets (Reichstein et al., 2019; Wang et al., 2022; Wang et al., 2023). In recent years, well-logging data-driven models have been successfully applied in various petrophysical problems, achieving excellent performance in terms of efficiency and accuracy (Sebtosheikh et al., 2015; Ao et al., 2019). Wang et al. (2012) employed multiclass neural networks to predict shale lithologies and organic-matter richness based on core analysis and conventional and neutron spectroscopy (PNS) logs across the Appalachian Basin. Kim et al. (2020) proposed a deep neural network to predict mineral compositions from integrated XRD results and conventional logs. Elkatatny et al. (2018) developed an artificial neural network (ANN) model to estimate the permeability of a heterogeneous reservoir from resistivity, bulk density, and neutron porosity logs. Miah et al. (2020) employed a multi-layer perception (MLP) model and kernel function-based least-squares support vector machine techniques for water saturation predictions from the log data.

The Illinois Basin - Decatur Project (IBDP) was a large-scale pilot geological carbon storage project that targeted a deep saline aquifer for CO_2 injection. Another

129 project, the Industrial CCS (ICCS), has also been implemented in the same site and is
130 currently actively injecting into the same formation. Rigorous and extensive
131 geophysical measurements were acquired for both projects, such as well-logging
132 techniques, high-fidelity induced seismicity monitoring, and reservoir numerical
133 simulation, for pre-injection formation characterization and post-injection CO_2
134 monitoring. These measurements provide valuable datasets for future CCS projects.
135 This paper applied five commonly used supervised ML and DL models, including the
136 ridge regression (RR) (Hoerl & Kennard, 1970), gradient boosting regression (GBR)
137 (Friedman, 2001), random forest (RF) (Breiman, 2001), support vector regression (SVR)
138 (Cortes & Vapnik, 1995; Schölkopf et al., 2000; Smola & Schölkopf, 2004), and ANN
139 (Schmidhuber, 2015), to analyze conventional logging data and PNX logging data for
140 formation mineralogy interpretation and CO_2 saturation estimation, respectively. We
141 optimized the hyperparameters of the four ML models and the DL model using the
142 simulated annealing (SA) algorithm (Kirkpatrick et al., 1983; Suman & Kumar, 2006)
143 and the grid search strategy, respectively, to achieve the best performance with respect
144 to the evaluation metrics of the root mean square error (RMSE) and coefficient of
145 determination (i.e., R^2 score). In addition, we ranked the importance of the input
146 features for the CO_2 saturation estimation in the optimized models using the
147 permutation importance algorithm (Fisher et al., 2018), which aims to find the most
148 influential features.

149 As global recognition of the pressing need to reduce carbon emissions grows, the
150 development and deployment of geological carbon storage projects are critical to

achieving this goal. The systematic analysis of well-logging data from the Illinois Basin provides novel insights into CCS and addresses a key knowledge gap in this field. The development of well-trained data-driven models enhances the interpretation of logging data and facilitates the efficient and reliable estimation of mineralogy, porosity, and CO_2 saturation. The results presented in this study demonstrate the promising application of ML and DL approaches in interpreting logging data from carbon storage sites, paving the way for more effective and efficient characterization, monitoring, and management in CCS projects.

2. Methodology

Figure 1 illustrates the dataset and methodology overview of this work. Particularly, **Section 2.1** provides an introduction to the study site, the Illinois Basin, and the dataset obtained from the two CCS projects. The dataset includes both conventional and PNX well-logging data. **Section 2.2** presents an overview of the five supervised ML and DL models that were used in this study, including the RR, RF, GBR, SVR, and ANN. **Section 2.3** provides details of the model training process, including the evaluation metrics, the SA algorithm for hyperparameter optimization, and the permutation importance algorithm for input feature importance ranking of the CO_2 estimation models.

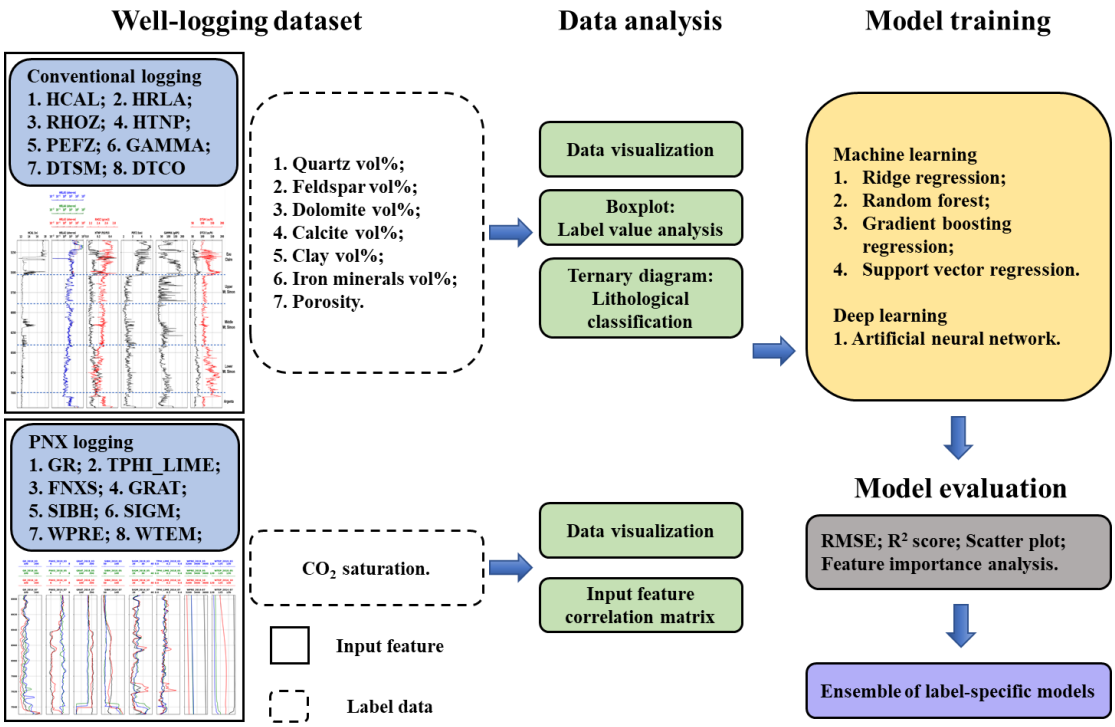


Figure 1. Dataset and Methodology overview.

2.1 Study site and dataset

The Illinois Basin is located in the central United States and covers an area of approximately 60,000 mile^2 across most of Illinois, southwest Indiana, and western Kentucky (Finley et al., 2014). The Mt. Simon Sandstone of the Illinois Basin is a saline aquifer that is present throughout the basin with thickness ranging from approximately 120 to 600 meters (400 to 2000 ft). This aquifer is ideal for CO_2 storage with adequate capacity and injectivity and is overlain by the impermeable Eau Claire Formation as a caprock. The IBDP and ICCS are two integrated bioenergy carbon capture and geological storage projects permanently storing CO_2 from ethanol production. Each project has one injection well (IBDP: CCS1, and ICCS: CCS2) and one deep

183 monitoring well (IBDP: VW1, and ICCS: VW2). The CO_2 injection of IBDP occurred
184 from November 2011 through November 2014, with a total injection amount of one
185 million tonnes of supercritical CO_2. The perforation locations were in one of the highest
186 permeability and porosity lithological units of the Lower Mt. Simon section, with a
187 depth from 2129 to 2138 m (6985 to 7015 ft). The ICCS was designed to inject one
188 million tonnes of supercritical CO_2 annually, with injection commencing in April 2017;
189 the perforation depths were from 2035 to 2054 m (6675 to 6740 ft). Numerous thin (a
190 few meters thick) mudstone baffles with very low permeability are present within the
191 Mt.Simon, which appear to be ubiquitous but discontinuous. One of these baffles is
192 present at a depth of 2092 m (6863 ft) with a porosity of 0.015 and vertical permeability
193 less than 0.01 mD, which is between the injection intervals of the IBDP and ICCS. This
194 baffle and others in the reservoir are interpreted to have inhibited vertical flow and to
195 have significantly limited the interactions between these two injections (Williams-
196 Stroud et al., 2020).

197 The dataset consists of the conventional and PNX well-logging data, which are for
198 pre-injection mineralogy interpretation and post-injection CO_2 saturation estimation,
199 respectively. The conventional well-loggings were collected from CCS1, CCS2, and
200 VW2, including 11,700 data points. The latest-generation pulsed neutron tool, PNX,
201 was time-lapse logged on CCS1 and VW1 between July 2014 and March 2019,
202 including 1600 data points. **Figure 2** illustrates the conventional well-logging curves
203 of CCS1, and **Figure 3** shows the boxplot of the interpreted volumetric fractions of 11
204 minerals and formation porosity (PHIT). The input features include 11 conventional

logging curves representing the caliper (HCAL), the high-resolution laterolog array for formation resistivity in four modes (HRLA2, HRLA3, HRLA4, and HRLA5), bulk density (RHOZ), high-resolution thermal neural porosity (HTNP), photoelectric absorption properties (PEFZ), total radioactivity (GAMMA), shear wave delay time (DTSM), and compressional wave delay time (DTCO). The data was obtained in an interval of 0.5 ft from about 5100 to 7200 ft, covering most Eau Claire formations (i.e., the caprock) and the entire Mt. Simon formation (i.e., the storage formation).

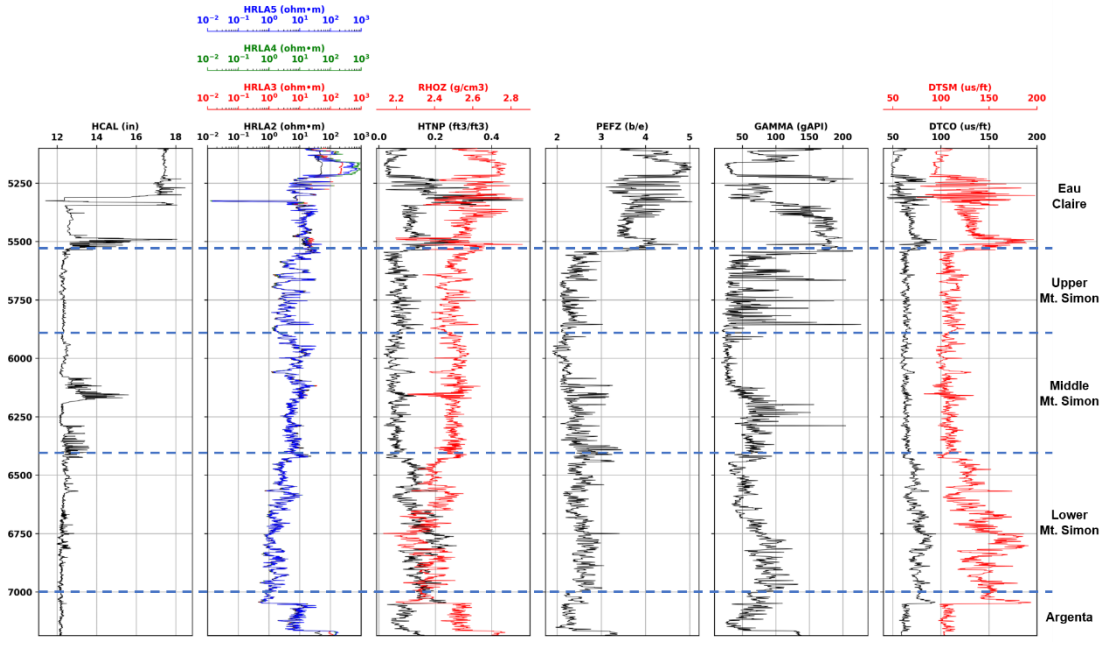
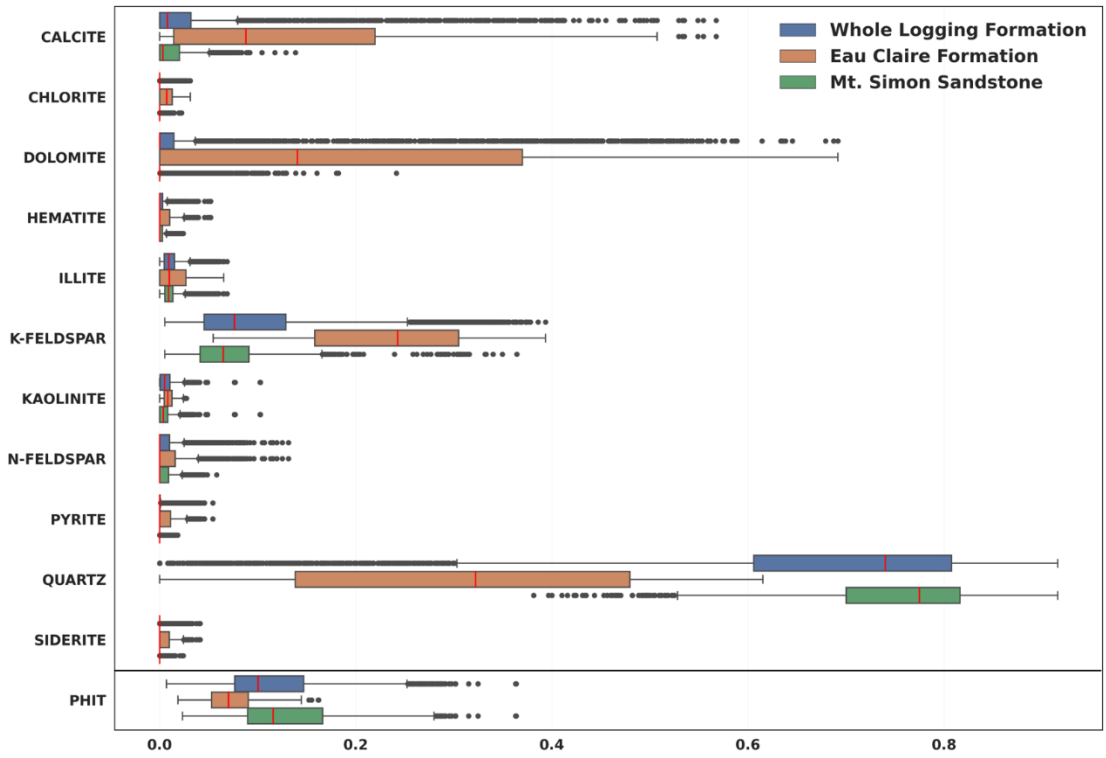


Figure 2. Conventional well-logging curves of CCS1 with the strata.



215

216 **Figure 3.** Boxplot of volumetric fractions of minerals and porosity in CCS1. The red
217 line in the box represents the median values of different minerals and porosity. Two
218 ends of the box indicate the 25th (Q1) and 75th (Q3) percentile, and the box length (i.e.,
219 Q3-Q1) is the interquartile range (IQR). The whiskers indicate values ranging from
220 0.35th to 99.65th percentile. The data out of whiskers are outliers.

221

222 Quartz is the dominant mineral within the conventional log data of CCS1 with an
223 average volume fraction of 66.3%. K-feldspar is the second with an average volume
224 fraction of 10.4%. N-feldspar, kaolinite, illite, dolomite, and calcite are the next in the
225 abundance with an average volume fraction of about 1% to 5%, followed by siderite,
226 pyrite, hematite, and chlorite with fractions less than 0.2%. The noticeable differences
227 in formation porosity and mineral abundance, especially the quartz, K-feldspar, calcite,

and dolomite, indicate the significant lithological variations in the Eau Claire Formation and Mt. Simon Sandstone. **Figure 4** shows the lithological classifications based on the mineral compositions using the ternary diagram (Glaser et al., 2013). The results indicated that the Eau Claire Formation consists primarily of fine-grained sandstone, siltstone, and minor dolomite with low porosity, which is about 0.072. The primary lithology in the Mt Simon Sandstone is coarse-grained and poorly sorted sandstone. Particularly, the average porosity of the injection target formation of the IBDP and ICCS, Lower Mt. Simon, is about 0.187, with a thickness of 600 ft. The minor minerals with similar properties are combined. The final label data includes porosity and volumetric fractions of quartz, feldspar (i.e., N-feldspar and K-feldspar), dolomite, calcite, total clay (i.e., kaolinite, illite, and chlorite), and iron minerals (i.e., siderite, pyrite, and hematite).

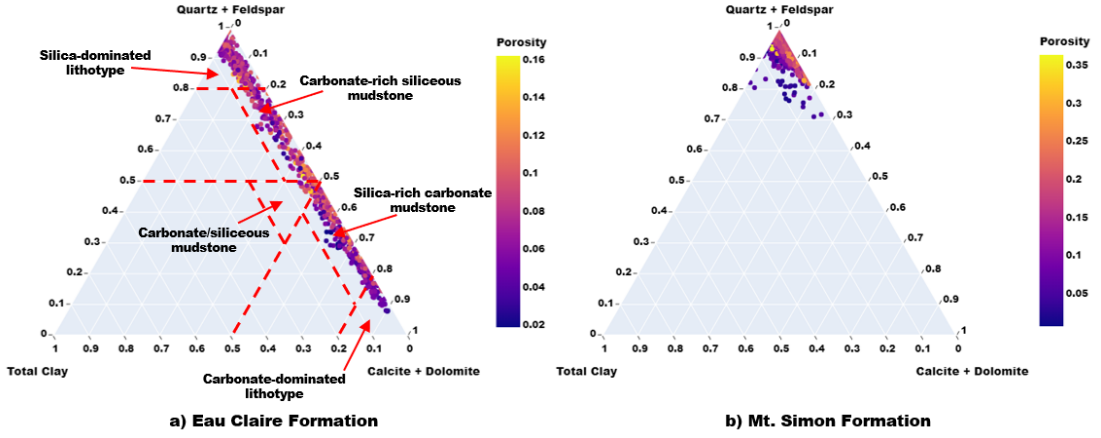


Figure 4. Lithological classification based on mineral compositions. The marker's color represents the porosity value.

The measurements of PNX well-loggings include the PNX sigma (GR), fast neutron cross-section (FNXS), gas ratio (GRAT), PNX sigma borehole corrected (SIBH), formation sigma (SIGM), PNX thermal neutron porosity hydrogen index (TPHI_LIME), well pressure (WPRE), and well temperature (WTEP). The PNX logging depth is from 6895 through 7045 ft with an interval of 0.5 ft. Particularly, the FNXS and GRAT, which are based on the gamma rays resulting from inelastic neutron interactions, are highly sensitive to CO_2 or other gases. **Figure 5** illustrates the PNX curves and the interpreted CO_2 saturation of CCS1 at different logging times. The CO_2 saturation in July 2014 and October 2016 were similar, higher than the values in May 2018 and March 2019 because of the migration and dissolution dynamics of the injected CO_2 plume. Correspondingly, the WPRE, WTEP, FNXS, GRAT, and SIGM had the most significant change at different logging times, followed by other PNX measurements. **Figure 6** demonstrates the correlation matrix between the input features and CO_2 saturation, computed using the Pearson correlation coefficients of the log data. Among all the input features, the SIBH and GRAT, with a correlation coefficient of -0.71, and WPRE and WTEP, with a correlation coefficient of 0.72, showed strong interdependence. In addition, it was observed that the label data of CO_2 saturation had no strong dependency on all the input features, thereby indicating the needs of using ML and DL models to explore the hidden correlations between data.

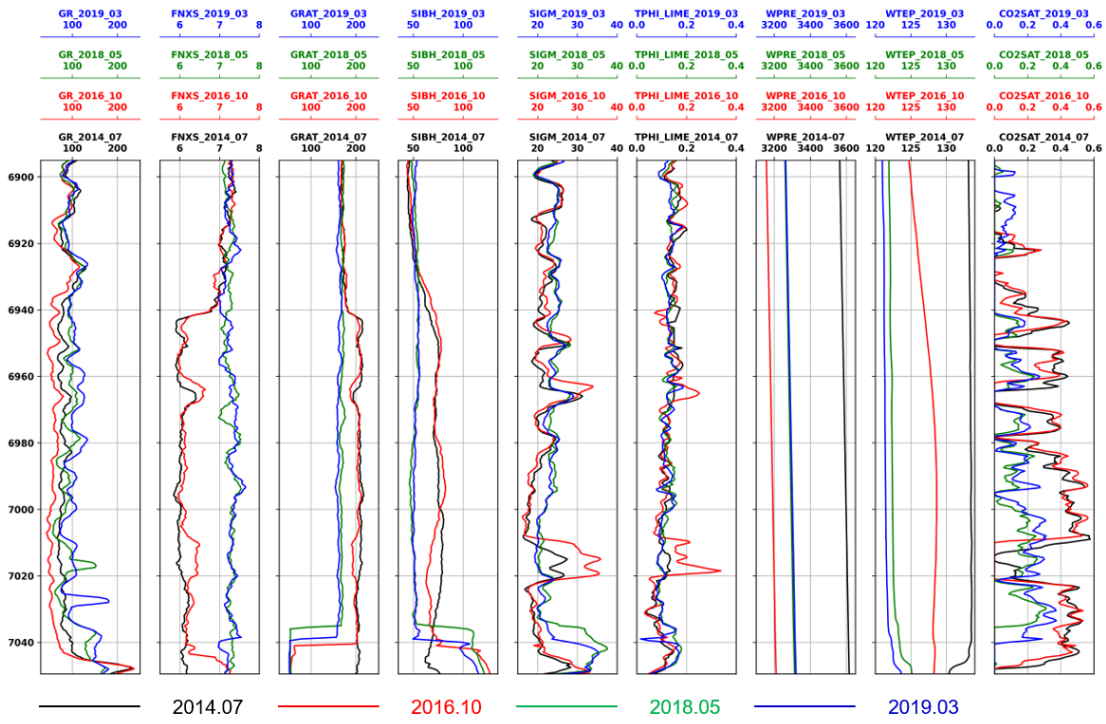


Figure 5. PNX well-logging measurement and CO_2 saturation curves of CCS1. The line's color represents the logging time.

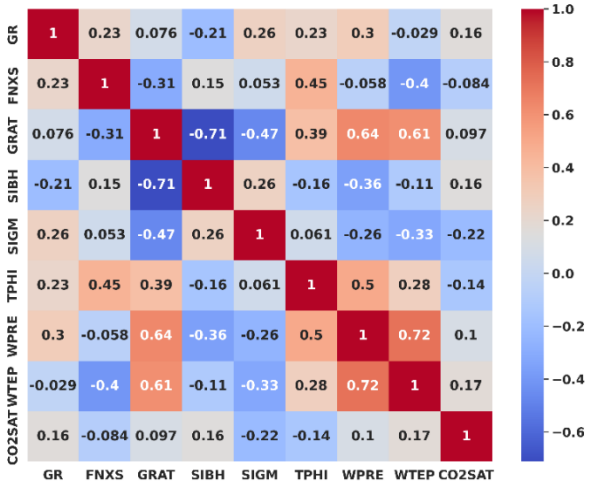


Figure 6. Correlation matrix of the PNX well-logging data.

2.2 Machine learning algorithms

2.2.1 Ridge regression

The RR (Hoerl & Kennard, 1970) is a modified form of linear regression that introduces an L2 penalty term to the least square loss function. The purpose of this additional term is to constrain the magnitude of the regression coefficients, which can help reduce the impact of multicollinearity and improve the stability and accuracy of the model (Duzan & Shariff, 2015). The model and the loss function can be written as:

$$y = xW + W_0 \tag{1}$$

$$L = \sum_{i=1}^{m} (Y_i - y_i)^2 + \alpha \sum_{j=1}^{n} W_j^2 \tag{2}$$

where $x \in R^{m \times n}$ represents the input feature matrix; $y \in R^m$ and $Y \in R^m$ are the predicted and true label data values, respectively; $W \in R^n$ and W_0 denote the vector of linear regression coefficients and the intercept, respectively.; α is the regularization parameter, which controls the trade-off between the magnitude of the coefficients and model generalization to prevent model overfitting. A larger value of α means greater shrinkage of the coefficients and results in a simpler and more stable model. In practice, the optimal α value is searched in the range between 0 and 10.

2.2.2 Random forest regression

The RF (Ho, 1995; Breiman, 2001) is an ensemble learning method that combines

multiple decision trees to construct a robust and accurate model for regression problems. **Figure 7** shows a diagram of the RF. The algorithm randomly selects a subset of the training data and a subset of the input features to build each decision tree, which helps reduce overfitting and improve generalization performance (Liaw & Wiener, 2002). The final prediction is obtained by averaging the predictions of all decision trees. The RF can handle high-dimensional data with complex interactions between features and non-linear relationships between input features and output variables; it is also less sensitive to missing data and outliers. In this study, we focused on two critical hyperparameters of the RF model: the number of decision trees in the forest and the number of features considered for the optimal split at each node. To determine the optimal values for these hyperparameters, we searched within specified ranges; the former was explored within the range of [10, 200], whereas the latter was investigated within the range between two and the maximum input feature number.

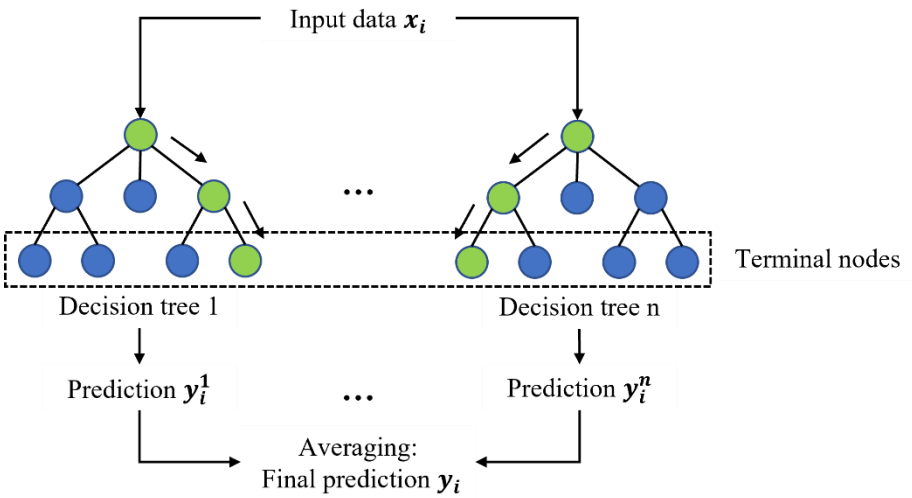


Figure 7. Random forest diagram for regression. Each decision tree produces a prediction at the terminal node that the input data belongs to, shown in the paths of the

green nodes and arrows. The terminal nodes are the last layer of decision trees shown in the dashed box.

2.2.3 Gradient boosting regression

The GBR (Friedman, 2001) builds an ensemble of weak prediction models (i.e., decision trees) and iteratively adds trees to the model in a forward step-wise manner. Each tree aims to improve the performance of the previous tree by fitting the residuals (Bentéjac et al., 2021), which are the differences between the predicted and true label data values:

$$y^t(x) = y^{t-1}(x) + v\sum_{jt=1}^{J_t}\gamma_{jt}\delta\left(x, R_{jt}\right) \quad (3)$$

where y^t is the updated prediction of model at iteration t, and y^{t-1} is the previous prediction at iteration $t-1$; v is the learning rate; R_{jt} denotes the *j*-th terminal node in the *t*-th decision tree (i.e., iteration); J_t is the total number of terminal nodes (i.e., leaves) in the *t*-th decision tree; $\delta(x,\ R_{jt})$ is equal to one if the given x falls in the terminal node R_{jt}, and zero otherwise; γ_{jt} is the optimal value that minimizes the loss function on each terminal node, written as:

$$\gamma_{jt} = \underset{\gamma}{\operatorname{argmin}}\sum_{x_i \epsilon R_{jt}} L(Y_i, y^{t-1}(x_i) + \gamma) \quad (4)$$

where x_i is the input that belongs to R_{jm}, and L is the loss function.

At each iteration, the algorithm calculates the negative gradient of the loss function

(e.g., RMSE) with respect to the current prediction. A weak model is then fitted to the negative gradient and included in the ensemble. The process is repeated until the loss function is minimized or a pre-specified stopping criterion is met. The key hyperparameters for the GBR model include the number of weak models (i.e., decision trees) to be included in the ensemble, which is searched within the range of [10, 200], and the learning rate, which controls the contribution of each weak model to the final prediction and is explored within the range of [0, 1].

2.2.4 Support vector regression

The SVR (Cortes & Vapnik, 1995) predicts continuous output values by constructing hyper-planes in the higher-dimensional space based on a subset of the training data (i.e., support vectors), which aims to control the deviation of the predicted value from the true value within a certain margin (Schölkopf et al., 2000; Smola & Schölkopf, 2004). To achieve this, the SVR uses a cost function that penalizes deviations from the true value and a regularization term that controls the complexity of the model. The cost function is based on the concept of ε-insensitive loss, where errors below a certain threshold (i.e., ε) are ignored:

$$L = \frac{1}{2}\sum_{j=1}^{n} W_j^2 + C\sum_{i=1}^{m} e_{\varepsilon i} \tag{5}$$

$$e_{\varepsilon i} = \begin{cases} 0 & if \quad |Y_i - y_i| \le \varepsilon \\ |Y_i - y_i| - \varepsilon & otherwise \end{cases} \tag{6}$$

where y_i is the predicted label value obtained using the linear hyperplanes, of which

the orientation and slope are determined by the weight vector, W; Y_i is the true label value; ε is the maximum allowable error around the hyperplane, and $e_{\varepsilon i}$ is the ε-intensitive loss, which penalizes errors greater than the value of the ε; C determines the trade-off between the model complexity and generalization.

In the case where the hyper-planes are non-linear, kernel functions are used to implicitly transform the input data into a higher-dimensional feature space, where the relationship between the input and output variables may become linear (Bagheripour et al., 2015). This effectively expands the feature space, making it possible to generate linear hyperplanes that can better capture the relationship between the input and output variables. The kernel functions calculate the similarity between pairs of input data points in the original feature space and represent this similarity as a dot product in the higher-dimensional feature space:

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \tag{7}$$

where $\varphi(x)$ is the projection of x in the transformed feature space. The hyperparameters of SVR are the kernel type, including linear, polynomial, radial basis function (RBF), and sigmoid, the regularization parameter C, and the ε parameter that specifies the margin of the tolerance around the regression hyperplanes. The optimal value of C is searched within the range of [0.1, 10], and the optimal value of ε is investigated within the range of [0.001, 0.1].

2.2.5 Artificial neural network

The MLP is one of the ANNs used for regression, which consists of layers (i.e., input, output, hidden) of interconnected nodes (i.e., neurons) performing non-linear transformation from input to output data (Schmidhuber, 2015). **Figure 8** shows the model architecture of a MLP network with two hidden layers.

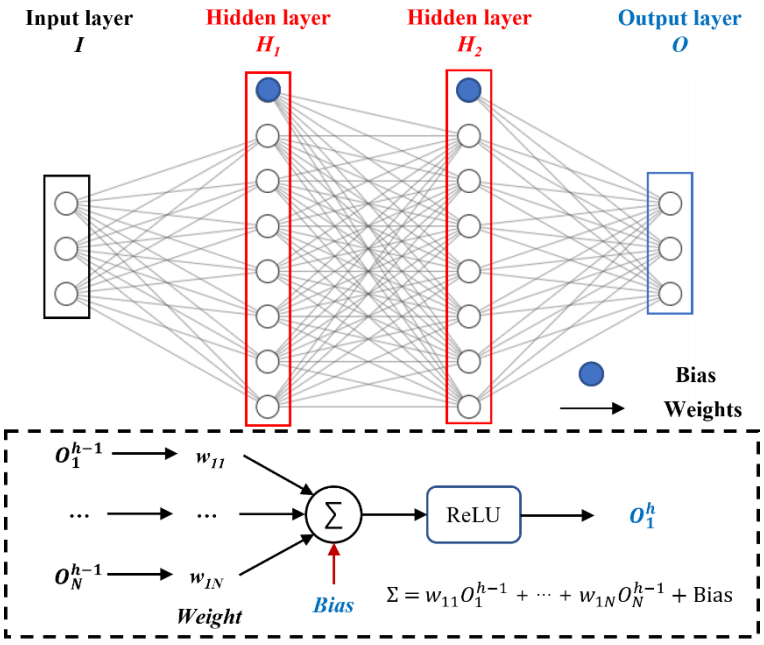


Figure 8. Model architecture of the MLP with two hidden layers. The dashed box provides the computation details at a hidden layer neuron.

Each node receives input from the previous layer and applied a linear combination of the input values, followed by a non-linear activation function that aims to capture potential nonlinear relations between input and output data. The output of the neuron is then passed to the next layer of neurons as input and the final layer of neurons produces the output of the model. The operation at each node can be written as:

$$O_i^h = \sigma_h\left(\sum_{j=1}^{N^{h-1}} \omega_{ij}^{h-1} O_j^{h-1} + b_i^{h-1}\right) \quad (8)$$

where O_i^h is the output of the node *i* at the layer *h*; σ_h is the activation function at the layer *h,* which is the rectified linear unit (ReLU) in this study; ω_{ij}^{h-1} is the weight between node *i* at layer *h* and the node *j* at the previous layer; b_i^{h-1} is the bias term, and N^{h-1} is the node number of layer *h-1*.

The MLP employs back-propagation (Rumelhart et al., 1986) with gradient descent to reduce the loss function (e.g., RMSE), which calculates the local gradient on the loss function surface and then updates the weight values along the maximum gradient direction with the learning rate. The back-propagation iteration includes a forward process for label data prediction and loss function calculation and a backward process for weight updating, which will be implemented repeatedly until the maximum iteration epoch or the pre-defined error tolerance is reached.

2.3 Model training and evaluation metrics

Prior to commencing the model training, input data underwent normalization by subtracting the mean value of each feature and scaling the values to achieve unit variance. The datasets, composed of normalized conventional and PNX well-logging data along with corresponding interpretation results (i.e., volumetric fractions of minerals, porosity, and CO_2 saturation), were divided into training and testing datasets at a ratio of 0.8 to 0.2. The ML (i.e., RR, RF, GBR, SVR) and DL (i.e., ANN) models

were then trained using the training dataset to identify patterns and relationships between the input and label data. The models' performance was evaluated using the testing dataset, which remained unseen during the training process. The evaluation metrics employed in this study were the RMSE and R^2 scores:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - y_i)^2} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Y_i - y_i)^2}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2} \quad (10)$$

where m is the number of data points; Y_i and y_i are the true and predicted label values, respectively; $\bar{Y}$ is the average value of Y_i.

The hyperparameters of the DL model were the number of hidden layers and number of neurons within each hidden layer. The number of hidden layers was examined within the range of 2-4, whereas the number of neurons per hidden layer was explored within the set of {50, 100, 150, 200}. A grid search strategy was employed to systematically evaluate various combinations of these values to identify the configuration that achieves the optimal performance. The training epoch and learning rate for the DL model were 200 and 0.001, respectively. The training and testing of the DL model (i.e., MLP) were implemented using Keras (Chollet et al., 2015; Abadi et al., 2016) with the TensorFlow backend on an NVIDIA Tesla T4 GPU.

The ML models (i.e., RR, RF, GBR, SVR) were trained and tested using the Scikit-learn package (Pedregosa et al., 2011). The hyperparameters of the four ML models

were optimized using a stochastic search optimization algorithm, SA (Kirkpatrick et al., 1983), which aimed to efficiently search the hyperparameter space and identify the best model configurations that yield superior performance in label data value prediction. The algorithm starts with an initial solution and then iteratively perturbs it to explore the search space. In each iteration, the perturbed solution is accepted if the objective function (e.g., RMSE) is improved, indicating a decrease in the loss function. If the perturbed solution results in a worse objective function value, it may still be accepted with a certain probability, which decreases as the difference between the current optimal and perturbed solutions increases. Additionally, the acceptance probability is governed by the temperature, which gradually decreases over time according to a predefined cooling schedule:

$$P(\Delta E) = e^{-\frac{\Delta E}{t}} \tag{11}$$

where P is the acceptance probability if $\Delta E > 0$, which means the perturbed solution causes higher energy (i.e., loss function) than current optimal solution; t is the temperature, which is reduced by the cooling parameter (i.e., $t^N = \alpha t^{N-1}$) after each iteration. The initial temperature is set at 100 and the cooling parameter, α, is set at 0.99. The stopping criteria are met when either the difference in the RMSE between two iterations is less than 10^{-4} or the maximum iteration number of 1000 is reached.

The feature importance ranking of the trained models for CO_2 saturation estimation was implemented using the permutation importance algorithm (Fisher et al., 2018), which calculates the change in the evaluation metrics (i.e., MSE or RMSE) output after

permuting the feature j in the input data D:

$$FI_j = MSE(Y, F(D)) - \frac{1}{K}\sum_{k=1}^{K} MSE\left(Y, F\left(D_j^k\right)\right) \quad (12)$$

where FI_j is the feature importance (FI) value of feature j; Y is the true label value; F is the trained model; K is the repetition number, which is 50 in this work; D is the original input data, and D_j^k is the input data with permutated feature j at the repetition k.

3. Results

3.1 Mineralogy interpretation

The goal of this study was to develop and evaluate various ML and DL models (i.e., RR, RF, GBR, SVR, and ANN) using features from 11 conventional well-logging inputs to predict seven label values, including the volumetric fractions of six minerals and the porosity, as shown in **Table 1**. For each ML algorithm (i.e., RR, RF, GBR, and SVR), we optimized and trained seven models to predict the seven individual data values. Table 1 presents the optimized hyperparameters for the various ML models. Additionally, a post-processing normalization step was applied to ensure that the sum of the seven predicted label values was equal to one. For the DL algorithm, because the single-output MLP models performed poorly with all possible hyperparameter combinations, one multi-output MLP model was optimized and trained. The loss function is the RMSE, and an additional constraint that the sum of all predicted label

values was equal to one was added. The best hyperparameters were two hidden layers, each of which contained 150 neutrons. All the models were trained using 9400 data points, each containing the 11 input features and the seven label values.

Table 1. Optimized hyperparameters of the ML models. The column μ indicates the average label value and σ indicates the standard deviation of the entire dataset. The coefficient of variation (CV) is the ratio of standard deviation to mean, representing the relative dispersion of data with respect to the mean. α is the regularization parameter in RR; n_e and n_f represent the number of estimators and features in RF, respectively; n_m and v indicate the number of weak models and learning rate in GBR, respectively; C and ε mean the regularization parameter and maximum tolerance in SVR, respectively. The values in the parentheses represent the gamma value of the RBF kernel function.

	μ	σ	CV	RR	RF		GBR		SVR		
				α	n_e	n_f	n_m	v	C	ε	Kernel
Quartz	0.702	0.168	0.239	0.033	150	7	98	0.205	4.75	0.017	RBF(0.39)
Feldspar	0.097	0.084	0.866	0.032	140	9	98	0.255	7.76	0.020	RBF(0.28)
Dolomite	0.027	0.083	3.704	0.019	93	9	99	0.246	15.95	0.019	RBF(0.31)

Calcite	0.017	0.048	2.824	0.010	75	8	100	0.218	2.40	0.003	RBF(0.95)
Clay	0.028	0.027	0.964	0.010	157	10	97	0.221	1.31	0.002	RBF(1.04)
Iron minerals	0.006	0.011	1.833	0.014	51	8	100	0.295	2.93	0.003	RBF(0.26)
Porosity	0.121	0.055	0.455	0.010	149	6	103	0.225	8.16	0.003	RBF(0.14)

The model performance based on the RMSE and R^2 score was tested on 2300 data points. **Table 2** presents the RMSE and R^2 scores of the ML and DL models in mineralogy interpretation. The overall evaluation metrics, including the RMSE and R^2 score, were calculated as the sum of each individual evaluation metric, weighted by the corresponding average label value. **Figure 9** displays the boxplot of the R^2 score of mineral and porosity predictions using the various models.

484 **Table 2.** Performance of model testing in mineralogy interpretation. The best
485 performances are shown in bold.

	RR		RF		GBR		SVR		ANN	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Quartz	0.063	0.862	**0.028**	**0.973**	0.028	0.972	0.030	0.967	0.036	0.955
Feldspar	0.029	0.879	**0.014**	**0.971**	0.015	0.967	0.018	0.953	0.020	0.944
Dolomite	0.049	0.648	0.022	0.927	**0.022**	**0.932**	0.026	0.899	0.028	0.885
Calcite	0.033	0.538	**0.017**	**0.883**	0.017	0.878	0.020	0.832	0.023	0.773
Clay	0.020	0.437	0.009	0.894	**0.005**	**0.964**	0.006	0.952	0.008	0.815
Iron minerals	0.006	0.721	0.003	0.939	**0.003**	**0.940**	0.003	0.913	0.006	0.699
Porosity	0.015	0.923	0.008	0.977	**0.008**	**0.980**	0.008	0.978	0.011	0.960
Overall	0.051	0.846	0.023	0.968	**0.023**	**0.969**	0.025	0.962	0.030	0.944

486

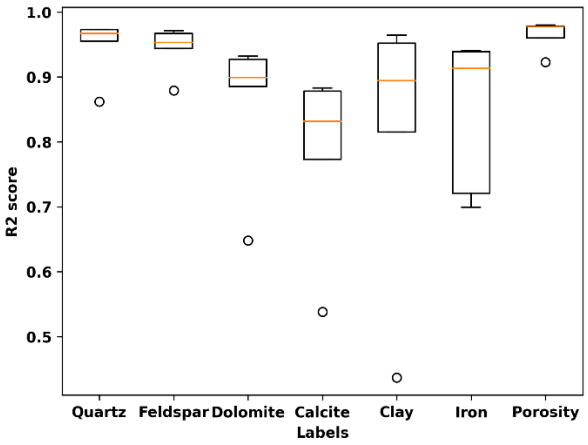


Figure 9 Boxplot of the R^2 score of mineral and porosity predictions.

The GBR and RF consistently outperformed the other three ML and DL models (i.e., RR, SVR, and ANN). Both the GBR and RF are ensemble learning methods that aggregate the predictions of multiple base learners, which helps reduce overfitting and improve generalization. The GBR, in particular, employs a gradient boosting algorithm that builds an ensemble of weak learners (i.e., decision trees) by iteratively optimizing a differentiable loss function and learning from residuals. This process allows GBR to be more adaptable and robust in various situations, leading to its lower sensitivity to hyperparameters within the conventional well-logging dataset. The boxplot analysis revealed that the labels with low CV values (porosity, quartz, and feldspar) generally had better prediction performance across different models. This observation can be attributed to the fact that low CV values indicate a lower degree of data dispersion, meaning that the data points are more closely clustered around the mean. When data is more homogeneously distributed, the underlying patterns and relationships are easier to capture for models, resulting in more accurate predictions. Conversely, labels with

504 higher CV values (clay, iron minerals, dolomite, and calcite) posed relatively greater
505 challenges in the prediction due to the larger dispersion of data points, which introduces
506 more complexity and variability, thereby making it more difficult for models to identify
507 and capture the underlying patterns. By using the highest-prediction-accuracy model
508 for each label (RF for quartz, feldspar, and calcite; GBR for the remaining labels), the
509 overall performance reached an R^2 score of 0.970 and an RMSE of 0.023, further
510 improving prediction accuracy. **Figure 10** presents a scatter plot to illustrate the true
511 and model-predicted label values from the best-performing models, which
512 demonstrates the effectiveness of the ensemble of label-specific models in capturing
513 the relationships between input features and label values.

514

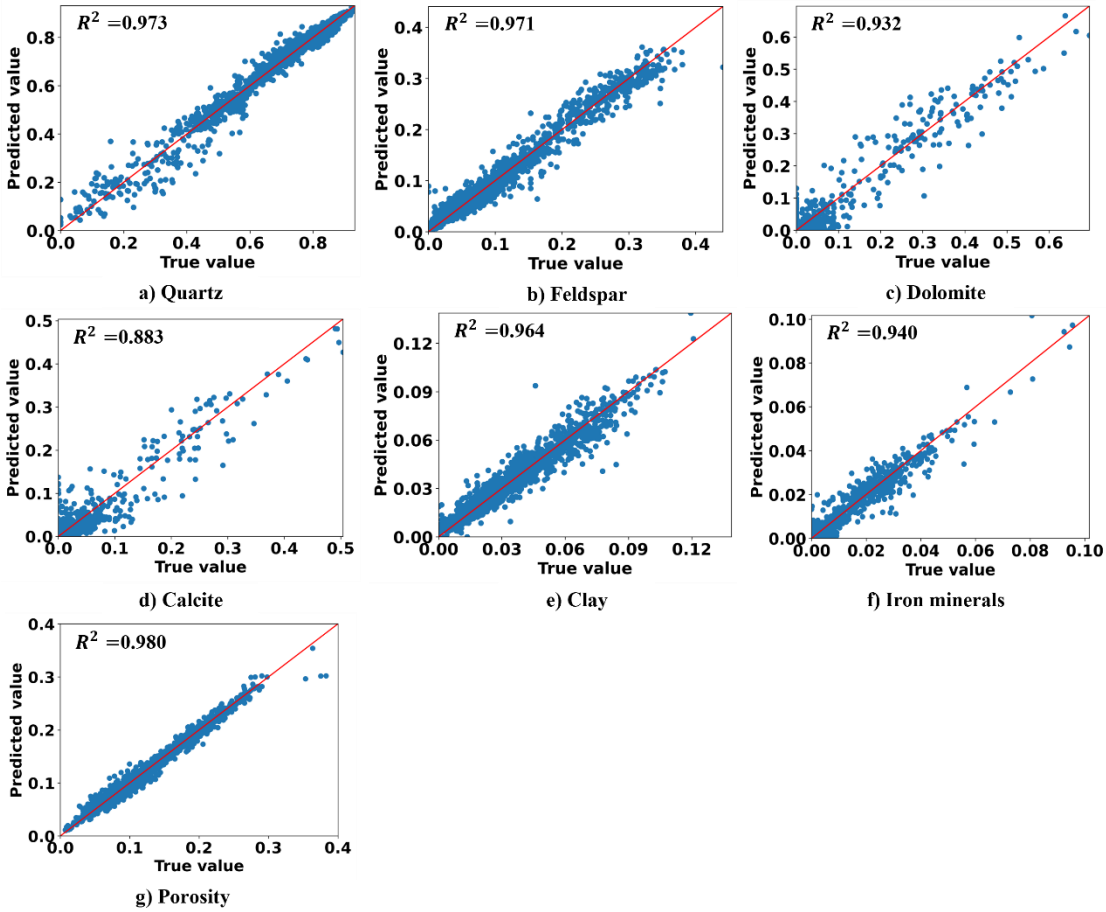


Figure 10. Scatter plots of mineralogy interpretation results of the best-performing models. The red lines are the references where true values are equal to predicted values.

3.2 CO_2 saturation estimation

We trained the various ML models (i.e., RR, RF, GBR, and SVR) and the DL model (i.e., MLP with two hidden layers and 150 neutrons each layer) using the PNX well-logging data, which included 1300 data points having eight input features and one label value (i.e., CO_2 saturation). The trained models were then tested on 330 data points. **Table 3** shows the optimized hyperparameters for the models, whereas **Table 4** illustrates the performance of model testing based on the RMSE and R^2 scores.

Table 3. Optimized hyperparameters of CO_2 saturation estimation models. The values in the brackets represent the gamma value of the RBF kernel function.

μ	σ	cv	RR	RF		GBR		SVR		
			α	n_e	n_f	n_e	v	C	ε	Kernel
0.137	0.162	1.182	0.033	198	4	105	0.234	8.58	0.015	RBF(1.04)

Table 4. Performance of model testing in CO_2 saturation estimation. The best performances are shown in bold.

RR		RF		GBR		SVR		ANN	
RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
0.101	0.640	**0.040**	**0.942**	0.050	0.910	0.041	0.940	0.048	0.920

Among the five ML and DL models, the RF exhibited the best performance in predicting the CO_2 saturation, achieving the highest R^2 score of 0.942. The ensemble nature of the RF model contributed to its robust performance by aggregating predictions from multiple decision trees. The SVR ranked the second with an R^2 score of 0.940,

followed by the ANN and GBR. The RR had the lowest R^2 score of 0.640. The high R^2 scores for the RF and SVR demonstrate their effectiveness in capturing complex non-linear relationships and interactions between the PNX well-logging data features and CO_2 saturation. **Figure 11** illustrates a scatter plot to compare the true and model-predicted label values.

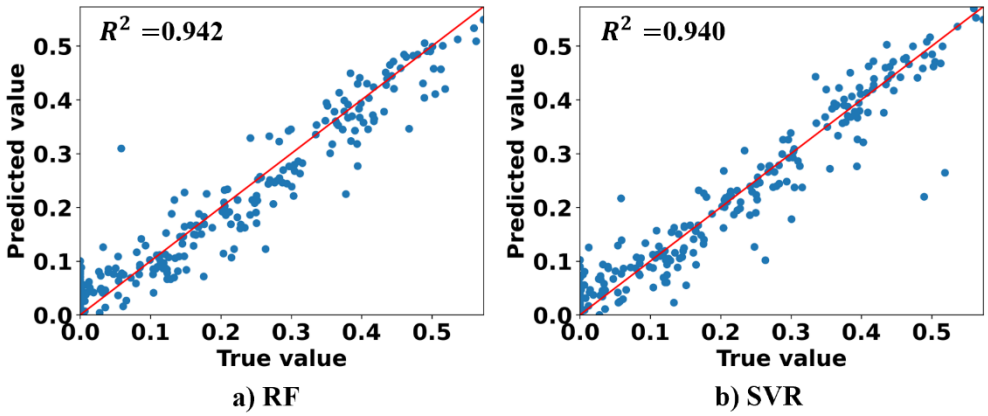


Figure 11. Scatter plot of CO_2 estimation results from the a) RF, and b) SVR models. The red lines are the references where true values are equal to predicted values.

Feature importance analysis was performed on the RF and SVR models using the permuted importance algorithm. **Figure 12** shows the FI values of the RF and SVR models. FI values measure the increase in model prediction error after feature values are permuted, thereby disrupting the relationships between input and label data. A large FI value indicates that the shuffled feature is more critical to model performance. The difference in feature importance rankings revealed the disparity in how each model captures the relationships between input features and CO_2 saturation. In the RF model,

the most important feature was SIGM, followed by FNXS, GRAT, and WPRE, which exhibited much lower feature importance. In contrast, the feature importance values in the SVR model were generally larger and more evenly distributed than those in the RF model, with WPRE and SIGM being the most important features. Despite the differences in feature importance, both the RF and SVR models identified SIGM and WPRE as significant contributors to the prediction of CO_2 saturation. This similarity suggests that these features can provide valuable information for predicting CO_2 saturation in the PNX well-logging dataset.

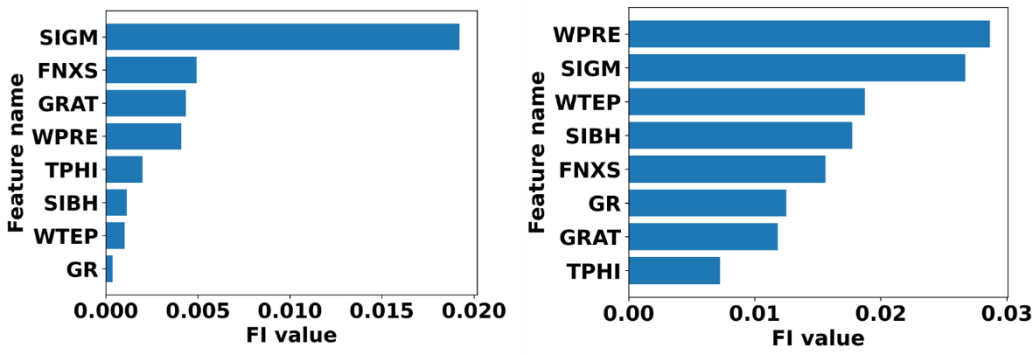


Figure 12. Feature importance analysis for the a) RF, and b) SVR models.

4. Conclusion

Carbon capture and storage (CCS) is a viable solution to mitigating continuous CO_2 buildup in the atmosphere and maintaining energy security under current technological and economical conditions. The geological carbon storage projects in the Illinois Basin represent the first large-scale CO_2 injection in saline aquifers in the United States. In

569 these projects, conventional well-logging was applied to evaluate the target formation
570 and the latest-generation pulsed neutron tool, PNX, was used to monitor the injected
571 CO_2. Subsurface saline aquifers are ideal storage sites due to their adequate capacity at
572 the global scale and relatively high injectivity. Therefore, the study on CCS in the
573 Illinois Basin will fill the knowledge gap in geological carbon storage and provide
574 valuable datasets for future CCS projects.

575 In this work, we developed and evaluated data-driven models based on five
576 commonly used ML and DL algorithms, namely the RR, RF, GBR, SVR, and ANN, for
577 formation mineralogy interpretation, based on 11 conventional well-logging input
578 features, and CO_2 saturation estimation, based on eight PNX well-logging input
579 features. The optimized hyperparameters were obtained using the SA algorithm for ML
580 models and the grid search strategy for the DL model. For mineralogy interpretation,
581 we trained seven single-output models for each ML algorithm and one multi-output
582 model for the DL algorithm to predict the volumetric fractions of six minerals and the
583 porosity. The optimized hyperparameters for most ML models vary significantly with
584 different labels, therefore highlighting the importance of optimizing individual models
585 for each label. Two ensemble learning methods, the GBR and RF, demonstrated
586 superior performance by achieving overall R^2 scores of 0.969 and 0.968, respectively.
587 Particularly, the GBR model showed relatively less sensitivity to hyperparameters for
588 the specific dataset used in this work. Additionally, label data with low CV values,
589 which included the porosity and volume fractions of quartz and feldspar, were more
590 likely to yield better performance across all models due to the low degree of label data

dispersion. The overall performance was further improved by strategically combining the best-performing models for each label, with the RF for quartz, feldspar, and calcite, and the GBR for the remaining labels. For CO_2 saturation estimation, the RF and SVR models exhibited the best performance, achieving R^2 scores of 0.942 and 0.940, respectively. The results of the feature importance analysis, conducted using the permutation importance algorithm, identified SIGM and WPRE as the most critical input features for CO_2 saturation estimation.

The ML and DL models capture the complex and non-linear relationships between input features and interpretation results (i.e., model predictions), thereby overcoming the limitations in time-consuming iteration methods and user-biased empirical models. The well-trained ML and DL models are efficient and reliable tools for well-logging data analysis and interpretation, which can be widely used in mineralogy interpretation for pre-injection formation evaluation and CO_2 saturation estimation for post-injection CO_2 monitoring in CCS projects.

Acknowledgement

The authors are thankful for the financial support provided by the University Coalition for Fossil Energy Research (UCFER) Program under the U.S. Department of Energy (DOE)'s National Energy Technology Laboratory (NETL) through the Award Number DE-FE0026825 and Subaward Number S000038-USDOE, as well as the support from NETL's Research Participation Program sponsored by the U.S. DOE and administered

by the Oak Ridge Institute for Science and Education (ORISE).

Reference

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. 2016 Mar 14.
2. Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174, 776-789.
3. Bachu, S. (2008). CO2 storage in geological media: Role, means, status and barriers to deployment. Progress in energy and combustion science, 34(2), 254-273.
4. Bagheripour, P., Gholami, A., Asoodeh, M., & Vaezzadeh-Asadi, M. (2015). Support vector regression based determination of shear wave velocity. Journal of Petroleum Science and Engineering, 125, 95-99.
5. Baumann, G., Henninges, J., & De Lucia, M. (2014). Monitoring of saturation changes and salt precipitation during CO2 injection using pulsed neutron-gamma logging at the Ketzin pilot site. International Journal of Greenhouse Gas Control, 28, 134-146.
6. Bauer, R. A., Carney, M., & Finley, R. J. (2016). Overview of microseismic response to CO2 injection into the Mt. Simon saline reservoir at the Illinois Basin-Decatur Project. International Journal of Greenhouse Gas Control, 54, 378-388.

633 7. Bauer, R., Will, R., E. Greenberg, S., & Whittaker, S. (2019). Illinois Basin–Decatur
634 Project. In T. Davis, M. Landrø, & M. Wilson (Eds.), Geophysics and
635 Geosequestration (pp. 339-370). Cambridge: Cambridge University Press.
636 doi:10.1017/9781316480724.020
637 8. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of
638 gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.
639 9. Boot-Handford, M. E., Abanades, J. C., Anthony, E. J., Blunt, M. J., Brandani, S.,
640 Mac Dowell, N., ... & Fennell, P. S. (2014). Carbon capture and storage
641 update. Energy & Environmental Science, 7(1), 130-189.
642 10. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
643 11. Cai, J., Wei, W., Hu, X., & Wood, D. A. (2017). Electrical conductivity models in
644 saturated porous media: A review. Earth-Science Reviews, 171, 419-433.
645 12. Chollet, F., & others. (2015). Keras. GitHub. Retrieved from
646 https://github.com/fchollet/keras
647 13. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3),
648 273-297.
649 14. Corina, A. N., & Hovda, S. (2018). Automatic lithology prediction from well
650 logging using kernel density estimation. Journal of Petroleum Science and
651 Engineering, 170, 664-674.
652 15. Clarkson, C. R., Vahedian, A., Ghanizadeh, A., & Song, C. (2019). A new low-
653 permeability reservoir core analysis method based on rate-transient analysis
654 theory. Fuel, 235, 1530-1543.

16. Dance, T., LaForce, T., Glubokovskikh, S., Ennis-King, J., & Pevzner, R. (2019). Illuminating the geology: Post-injection reservoir characterization of the CO2CRC Otway site. International Journal of Greenhouse Gas Control, 86, 146-157.
17. Duzan, H., & Shariff, N. S. B. M. (2015). Ridge regression for solving the multicollinearity problem: review of methods and models. Journal of Applied Science.
18. Elkatatny, S., Mahmoud, M., Tariq, Z., & Abdulraheem, A. (2018). New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial intelligence network. Neural Computing and Applications, 30(9), 2673-2683.
19. Ershadnia, R., Hajirezaie, S., Amooie, A., Wallace, C. D., Gershenzon, N. I., Hosseini, S. A., ... & Soltanian, M. R. (2021). CO2 geological sequestration in multiscale heterogeneous aquifers: Effects of heterogeneity, connectivity, impurity, and hysteresis. Advances in Water Resources, 151, 103895.
20. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
21. Finley, R. J. (2014). An overview of the Illinois Basin–Decatur project. Greenhouse Gases: Science and Technology, 4(5), 571-579.
22. Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res., 20(177), 1-81.
23. Gaus, I. (2010). Role and impact of CO2–rock interactions during CO2 storage in

677 sedimentary rocks. International journal of greenhouse gas control, 4(1), 73-89.
678 24. Glaser, K. S., Miller, C. K., Johnson, G. M., Toelle, B., Kleinberg, R. L., Miller, P.,
679 & Pennington, W. D. (2013). Seeking the sweet spot: Reservoir and completion
680 quality in organic shales. Oilfield Review, 25(4), 16-29.
681 25. Guo, R., Dalton, L. E., Fan, M., McClure, J., Zeng, L., Crandall, D., & Chen, C.
682 (2020). The role of the spatial heterogeneity and correlation length of surface
683 wettability on two-phase flow in a CO2-water-rock system. Advances in Water
684 Resources, 146, 103763.
685 26. Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd
686 international conference on document analysis and recognition (Vol. 1, pp. 278-
687 282). IEEE.
688 27. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for
689 nonorthogonal problems. Technometrics, 12(1), 55-67.
690 28. Holechek, J. L., Geli, H. M., Sawalhah, M. N., & Valdez, R. (2022). A global
691 assessment: can renewable energy replace fossil fuels by
692 2050?. Sustainability, 14(8), 4792.
693 29. Houseman, G. A., Cull, J. P., Muir, P. M., & Paterson, H. L. (1989). Geothermal
694 signatures and uranium ore deposits on the Stuart Shelf of South
695 Australia. Geophysics, 54(2), 158-170.
696 30. IEA (2022), Global Energy Review: CO2 Emissions in 2021, IEA, Paris
697 https://www.iea.org/reports/global-energy-review-co2-emissions-in-2021-2
698 31. IEA (2023), CO2 Emissions in 2022, IEA, Paris https://www.iea.org/reports/co2-

699 emissions-in-2022, License: CC BY 4.0
700 32. Juanes, R., Spiteri, E. J., Orr Jr, F. M., & Blunt, M. J. (2006). Impact of relative
701 permeability hysteresis on geological CO2 storage. Water resources
702 research, 42(12).
703 33. Jia, B., Tsau, J. S., & Barati, R. (2019). A review of the current progress of CO2
704 injection EOR and carbon storage in shale oil reservoirs. Fuel, 236, 404-427.
705 34. Kamali, M. R., & Mirshady, A. A. (2004). Total organic carbon content determined
706 from well logs using ΔLogR and Neuro Fuzzy techniques. Journal of Petroleum
707 Science and Engineering, 45(3-4), 141-148.
708 35. Kim, D., Choi, J., Kim, D., & Byun, J. (2020). Predicting mineralogy by integrating
709 core and well log data using a deep neural network. Journal of Petroleum Science
710 and Engineering, 195, 107838.
711 36. Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated
712 annealing. science, 220(4598), 671-680.
713 37. Lai, J., Wang, G., Wang, S., Cao, J., Li, M., Pang, X., ... & Qin, Z. (2018). Review
714 of diagenetic facies in tight sandstones: Diagenesis, diagenetic minerals, and
715 prediction via well logs. Earth-Science Reviews, 185, 234-258.
716 38. Leung, D. Y., Caramanna, G., & Maroto-Valer, M. M. (2014). An overview of
717 current status of carbon dioxide capture and storage technologies. Renewable and
718 Sustainable Energy Reviews, 39, 426-443.
719 39. Li, Z., Ripepi, N., & Chen, C. (2020). Using pressure pulse decay experiments and
720 a novel multi-physics shale transport model to study the role of Klinkenberg effect

721 and effective stress on the apparent permeability of shales. Journal of Petroleum
722 Science and Engineering, 189, 107010.
723 40. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R
724 news, 2(3), 18-22.
725 41. Lindberg, D. V., Rimstad, E., & Omre, H. (2015). Inversion of well logs into facies
726 accounting for spatial dependencies and convolution effects. Journal of Petroleum
727 Science and Engineering, 134, 237-246.
728 42. Lynas, M., Houlton, B. Z., & Perry, S. (2021). Greater than 99% consensus on
729 human caused climate change in the peer-reviewed scientific
730 literature. Environmental Research Letters, 16(11), 114005.
731 43. Markewitz, P., Kuckshinrichs, W., Leitner, W., Linssen, J., Zapp, P., Bongartz, R., ...
732 & Müller, T. E. (2012). Worldwide innovations in the development of carbon
733 capture technologies and the utilization of CO 2. Energy & environmental
734 science, 5(6), 7281-7305.
735 44. Miah, M. I., Zendehboudi, S., & Ahmed, S. (2020). Log data-driven model and
736 feature ranking for water saturation prediction using machine learning
737 approach. Journal of Petroleum Science and Engineering, 194, 107291.
738 45. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ...
739 & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of
740 machine Learning research, 12, 2825-2830.
741 46. Pires, J. C. M., Martins, F. G., Alvim-Ferraz, M. C. M., & Simões, M. (2011).
742 Recent developments on carbon capture and storage: An overview. Chemical

743 engineering research and design, 89(9), 1446-1460.
744 47. Renforth, P., & Henderson, G. (2017). Assessing ocean alkalinity for carbon
745 sequestration. Reviews of Geophysics, 55(3), 636-674.
746 48. Ren, X., Hou, J., Song, S., Liu, Y., Chen, D., Wang, X., & Dou, L. (2019). Lithology
747 identification using well logs: A method by integrating artificial neural networks
748 and sedimentary patterns. Journal of Petroleum Science and Engineering, 182,
749 106336.
750 49. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais,
751 N. (2019). Deep learning and process understanding for data-driven Earth system
752 science. Nature, 566(7743), 195-204.
753 50. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations
754 by back-propagating errors. nature, 323(6088), 533-536.
755 51. Rose, D., Zhou, T., Beekman, S., Quinlan, T., Delgadillo, M., Gonzalez, G., ... &
756 Bailly, J. P. (2015, July). An innovative slim pulsed neutron logging tool.
757 In SPWLA 56th Annual Logging Symposium. OnePetro.
758 52. Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New
759 support vector algorithms. Neural computation, 12(5), 1207-1245.
760 53. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural
761 networks, 61, 85-117.
762 54. Sebtosheikh, M. A., & Salehi, A. (2015). Lithology prediction by support vector
763 classifiers using inverted seismic attributes data and petrophysical logs as a new
764 approach and investigation of training data set size effect on its performance in a

765 heterogeneous carbonate reservoir. Journal of Petroleum Science and
766 Engineering, 134, 143-149.
767 55. Silversides, K., Melkumyan, A., Wyman, D., & Hatherly, P. (2015). Automated
768 recognition of stratigraphic marker shales from geophysical logs in iron ore
769 deposits. Computers & Geosciences, 77, 118-125.
770 56. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector
771 regression. Statistics and computing, 14(3), 199-222.
772 57. Suman, B., & Kumar, P. (2006). A survey of simulated annealing as a tool for single
773 and multiobjective optimization. Journal of the operational research society, 57,
774 1143-1160.
775 58. Szabó, N. P., & Dobróka, M. (2020). Interval inversion as innovative well log
776 interpretation tool for evaluating organic-rich shale formations. Journal of
777 Petroleum Science and Engineering, 186, 106696.
778 59. Tan, M., Song, X., Yang, X., & Wu, Q. (2015). Support-vector-regression machine
779 technology for total organic carbon content prediction from wireline logs in organic
780 shale: A comparative study. Journal of Natural Gas Science and Engineering, 26,
781 792-802.
782 60. Wang, G., & Carr, T. R. (2012). Marcellus shale lithofacies prediction by multiclass
783 neural network classification in the Appalachian Basin. Mathematical
784 Geosciences, 44(8), 975-1004.
785 61. Wang, H., Dalton, L., Fan, M., Guo, R., McClure, J., Crandall, D., & Chen, C.
786 (2022). Deep-learning-based workflow for boundary and small target segmentation

787 in digital rock images using UNet++ and IK-EBM. Journal of Petroleum Science
788 and Engineering, 215, 110596.
789 62. Wang, H., Dalton, L., Guo, R., McClure, J., Crandall, D., & Chen, C. (2023).
790 Application of unsupervised deep learning to image segmentation and in-situ
791 contact angle measurements in a CO2-water-rock system. Advances in Water
792 Resources, 104385.
793 63. Williams-Stroud, S., Bauer, R., Leetaru, H., Oye, V., Stanek, F., Greenberg, S., &
794 Langet, N. (2020). Analysis of microseismicity and reactivated fault size to assess
795 the potential for felt events by CO 2 injection in the Illinois Basin. Bulletin of the
796 Seismological Society of America, 110(5), 2188-2204.
797 64. Yang, Y., & Aplin, A. C. (2010). A permeability–porosity relationship for
798 mudstones. Marine and Petroleum Geology, 27(8), 1692-1697.
799 65. York, R., & Bell, S. E. (2019). Energy transitions or additions? Why a transition
800 from fossil fuels requires more than the growth of renewable energy. Energy
801 Research & Social Science, 51, 40-43.
802 66. Zastempowski, M. (2023). Analysis and modeling of innovation factors to replace
803 fossil fuels with renewable energy sources-Evidence from European Union
804 enterprises. Renewable and Sustainable Energy Reviews, 178, 113262.
805 67. Zhang, J. (2011). Pore pressure prediction from well logs: Methods, modifications,
806 and new approaches. Earth-Science Reviews, 108(1-2), 50-63.