Effectiveness of Machine Learning in Carbon Capture and Storage Site Selection

Netra Karthigeyan

Received June 06, 2025
Accepted September 03, 2025
Electronic access October 15, 2025

Carbon Capture and Storage (CCS) is an increasingly essential tool to address global carbon dioxide emissions, with the potential to reduce CO_2 emissions by up to 15% by 2050, playing a critical role in achieving the goals of the Paris Agreement. However, choosing the right sites for storage is a complex challenge. Previous studies have focused on geological suitability, excluding vital factors such as cost, power produced, emissions proximity, and transport access. This study takes an integrated, data-driven approach to CCS site selection. By combining global datasets with machine learning models and optimization tools, we have predicted regions which are viable for CO_2 storage. Rather than relying on location-specific assessments, this research has explored how predictive models can recognize patterns across regions and identify sites that balance environmental, logistical, and financial feasibility. We have also explored how the removal or inclusion of variables affects the accuracy of site prediction, reinforcing the importance of a multidimensional approach. The outcomes of this project offer a blueprint for CCS planning and enable collaboration between governments, policymakers, industries, and researchers. Ultimately, this research adds to the expanding field of climate solutions that rely on computational tools. Our findings validate the effectiveness of ML, particularly ANNs, in CCS site selection.

Keywords: Carbon Capture and Storage (CCS), Machine Learning (ML), site selection, Geographic Information System (GIS), Mixed Integer Linear Programming (MILP), data-driven research, climate change

Effectiveness of Machine Learning in Carbon Capture and Storage Site Selection

Carbon dioxide emissions have been rising rapidly, primarily due to anthropogenic activities which are activities caused by humans, especially those that impact the environment. The burning of fossil fuels, industrial production, and deforestation are examples of such activities which lead to severe environmental consequences like global warming and ocean acidification[1]. In 2024, CO_2 emissions from fossil fuels and land-use changes reached a record high of 37.8 billion metric tons which was a 0.8% increase from the previous year[2]. As the urgency to mitigate climate change escalates, reducing CO_2 emissions alone is insufficient and hence carbon capture and storage (CCS) has emerged as a crucial technology to actively remove and sequester CO_2 from the atmosphere.

As shown in Figure 1, CCS works by capturing CO_2 from industrial sources and storing it in rock formations such as depleted reservoirs and saline aquifers over 2,600 feet underground to prevent it from entering the atmosphere[3]. The effectiveness of CSS is highly dependent on the selection of optimal storage sites, which must be evaluated based on geological suitability, proximity to CO_2-emitting sources, transportation feasibility, and overall cost-efficiency.

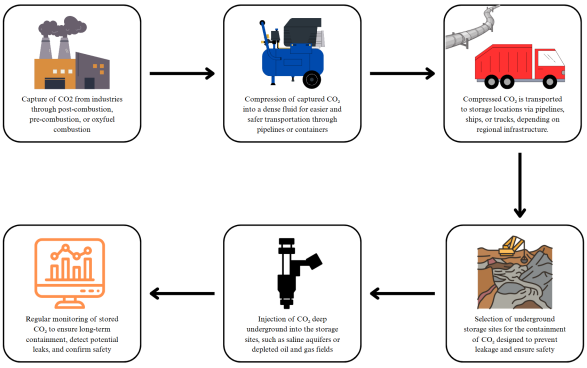

Fig. 1 A Simple Diagram Depicting the Carbon Capture and Storage Process

Carbon capture and storage (CCS) is becoming increasingly important as countries work toward ambitious climate goals such as Switzerland's 2050 net-zero target to ensure that greenhouse gas emissions do not exceed the amount that can be captured and stored in sinks[4]. Despite international agreements such as the Paris Accord, which aims to limit the temperature increase to 1.5°C above pre-industrial levels[5], current trajectories suggest a global temperature rise of 2.4 to 2.7°C, highlighting a major gap between commitments and action[6]. While CCS offers a way to

reduce atmospheric CO_2 levels, its large-scale success depends on selecting practical storage sites.

Existing research has made progress in modeling and planning for carbon capture and storage (CCS) using different tools. For example, one study applied artificial neural networks (ANNs) and support vector machines (SVMs) to optimize post-combustion carbon capture using monoethanolamine (MEA), a solvent that temporarily binds with CO_2, to enable separation and reuse [7]. The researchers demonstrated that ANNs, with their ability to learn nonlinear patterns in large datasets, outperformed SVMs in predicting CO_2 capture efficiency, showing the value of AI in improving carbon neutrality.

In a different context, researchers in the Netherlands used Geographic Information System (GIS)-based systems like ArcGIS and the MARKAL-NL-UU optimization model to design a cost-effective national CCS infrastructure [8]. They mapped sources and sinks of CO_2, planned primary routes, and linked spatial and economic data to optimize pipeline networks. The study found that integrating ArcGIS and MARKAL can guide long-term, policy-aligned infrastructure planning, though it also highlighted challenges such as high upfront costs and storage limitations in the future.

However, these models often operate in isolation, without combining geological, logistical, and financial factors into a single system. This research addresses that limitation by integrating machine learning, GIS, and mixed integer linear programming (MILP) into a single decision-making pipeline. The study aims to address the effectiveness of different machine learning algorithms to model optimal geographical site selection for CCS, considering the variables of geological suitability, proximity to a facility that emits CO_2, and cost-effectiveness. By using AI-driven models to enhance CCS efficiency, we can take a significant step toward reducing CO_2 levels in the atmosphere and mitigating the effects of climate change.

1 Methods

This research involved Exploratory Data Analysis (EDA) using machine learning to identify optimal sites for carbon capture and storage (CCS). The study focused on datasets on regions across the globe, such as the dataset curated by Banachewicz [9], that cover geological features, CO_2 emissions, and estimated CCS costs. The datasets were obtained from international organizations such as the International Energy Agency (IEA) and open-access platforms.

After the data was collected, it was pre-processed to resolve inconsistencies and format the datasets for effective analysis. The dataset contained categorical variables such as country location and continent name which were explicitly converted to indicator variables with Pandas. In addition, the dataset was normalized in order to have all numerical values in consistent units. This step was crucial for improving the accuracy of the models used. Machine learning played a key role in the predictive side of the project. Categorical and quantitative variables were processed within a pandas Data Frame to enable efficient manipulation and analysis of the dataset. The goal was to build models that can make accurate predictions on CCS site feasibility and power generated, which was evaluated by splitting the datasets into testing and training sets randomly, with 70% of the data used for training and 30% of the data used for testing, and subsequently checking for minimal values of mean squared error (MSE). The study did not employ cross validation.

This study primarily evaluates each model using its root mean squared error (RMSE) values. However, to validate the ANN's effectiveness, other metrics were also utilized. The model's Mean Absolute Error (MAE), which is the average magnitude of absolute differences between the predicted and actual values, was calculated. The R^2 statistic was also computed for the model. R^2, which should be 1 for an ideal model, measures how well the model's prediction align with the actual values and how well the independent variables or input features explain the dependent variable.

Geographic Information System (GIS) tools were used alongside this to analyze spatial data and assess location viability. Additionally, Mixed Integer Linear Programming (MILP) was utilized to ensure cost-efficiency site selection.

Although this study relies on publicly available and licensed datasets, ethical considerations remain essential. All sources were carefully examined for accuracy, potential bias, and responsible usage, and no confidential or personally identifiable information was accessed. The dataset used has been provided by an agency of the United States Government which has been published by the National Energy Technology Laboratory on their Carbon Capture and Storage Database, and is an anonymized dataset.

1.1 Machine Learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that allows computers to learn from data and make decisions or predictions. While AI is a broader concept that includes ML, ML itself is divided into three main types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning uses labeled data to predict outcomes, while unsupervised learning works with data that is not categorized and discovers hidden patterns [10]. Reinforcement learning involves an autonomous agent that make decisions and interact with an environment based on reward signals [11]. ML models improve their accuracy with more experience and datasets to train with [12]. These models are useful in predicting outcomes and classifying data, with minimal human intervention, thus enabling a simpler and more personalized process. ML played a key role in this project by helping analyze large global datasets and predicting site viability for Carbon Capture and Storage

(CCS).

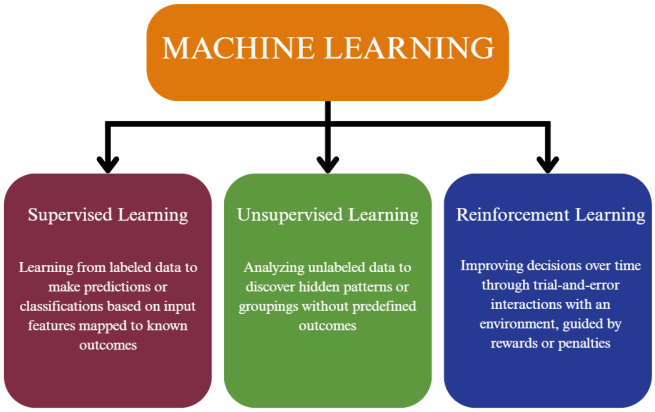

Fig. 2 An Overview of the Three Types of Machine Learnings

Machine Learning models can be black boxes, that is, the internal mechanism and factors affecting the result are unknown. In order to obtain the weight or contribution of each variable in the prediction of this study, the SHAP (SHapley Additive exPlanations) method was utilized. A positive SHAP value indicates positive impact on the prediction and while a negative value indicates negative impact. The magnitude of the SHAP value suggests the degree of impact the variable has on the result.

This research investigated the performance of ANNs for predicting CCS site viability in comparison with other ML algorithms – specifically, Support Vector Machines and decision trees – to assess its relative effectiveness.

1.2 Decision Trees

Decision Tree is a supervised machine learning algorithm used to predict continuous values. Decision Trees involve splitting the dataset based on its features, thus creating a tree-like structure where internal nodes indicate a point of decision, branches show the flow between decisions, and leaf nodes represent the point of final prediction [13]. Their performance can be adjusted using hyperparameters such as maximum depth, which limits the number of splits and helps prevent overfitting, and criterion, which is a function that measures the quality of the split. Research by Yan et al. [14], which discusses the role of ML in Carbon Capture, Utilization, and Storage (CCUS), acknowledges that while decision trees explain the logic of the model and reduce the black-box nature of ML models, they can lead to a decrease in predictive accuracy. In this study, we implemented a decision tree through the Scikit-learn library in Python with a maximum depth of 6 and plotted the tree to visualize its structure.

1.3 Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the best possible 'line' using kernels, to separate data into different classes. SVMs are divided into linear, radial, and polynomial machines based on the kernel each type employs. The performance of SVMs can be altered using hyperparameters such as different kernels; regularization (C), which controls the margin width, thus affecting the extent of generalization [15]; and gamma values, which determines the influence of a single training example [16]. In this project, we used SVM through the Scikit-learn library in Python, which provides simple tools to train and evaluate SVM models. Since SVM would have been useful in training the model to recognize patterns between variables, this research also investigated the effectiveness of ANNs for predicting CCS site viability compared to SVMs.

1.4 Neural Networks

Neural networks are a type of machine learning model that are made up of layers of connected nodes, also called neurons. The input layer receives the features or data, such as distance to facility or storage cost, in this research, and this data is passed through one or more hidden layers, ultimately passing through the output layer [12]. Each connection has a weight (used to determine the importance of each variable with respect to the output), that is adjusted during training to reduce error. Neural networks are trained over several cycles, called epochs, where the model makes a prediction, its performance is measured (usually using metrics like Mean Squared Error (MSE), which tells us how close the predictions are to the actual values), and then the weights are readjusted to reduce the error [17]. These models are powerful and can handle complex relationships in data – a feature that proved to be integral in this research.

The multilayer perceptron (MLP), the most commonly used type of neural network, is characterized by a feedforward architecture where signals generally move unidirectionally from input to output without loops. Its computational power comes from its non-linear activation functions which allow it to approximate any continuous function and act as a universal approximator [18]. The mathematical proof by Hornik et al. [19] demonstrates that the universal approximation ability of multilayer feedforward networks is not coincidental but rather reflects their inherent capacity to model complex functional relationships.

A study by Sipöcz et al. [20] employed a multilayer perceptron, for static regression applications in CO_2 capture plants. Cybenko [21] demonstrated that continuous feedforward neural networks with just one hidden layer and any continuous sigmoidal activation function, which is a mathematical function that maps input values to outputs between 0 and 1, are capable of approximating arbitrary decision boundaries with high accuracy. Accordingly, the research by Sipöcz et al. consisted of an input

layer, one or more hidden layers, and one output layer. The network was trained using backpropagation, which updates weights and biases based on the gradient of the error function, and was optimized using two algorithms - Levenberg–Marquardt (LM) and Scaled Conjugate Gradient (SCG). Notably, the LM-trained network achieved a maximum error of just 0.18%, outperforming the SCG-trained network, which showed higher prediction errors up to 1.4%. These findings highlight the effectiveness of MLPs for regression and the potential accuracy benefits of advanced optimization methods to significantly reduce the error. This prior research encouraged the selection of a multilayer perceptron to improve CCS site prediction accuracy and study the effect of change in the number of layers and nodes on the predictions.

1.5 Mixed Integer Linear Programming (MILP)

Mixed Integer Linear Programming (MILP) is a mathematical algorithm used to find the best solution to a problem while considering certain rules or limits. MILP handles integer variables rather than continuous values, and uses binary variables (zeroes and ones) to aid in effective decision making[22]. MILP has been used in used for cost optimization in carbon capture and storage, as seen in a study by Zhang et al.[23] which utilized the algorithm to optimize costs for the infrastructure of CO_2 transport networks.

MILP helps choose the most suitable sites and routes by balancing variables such as cost, transport, and geological factors, thus offering a practical output. We plan to use the SciPy library in Python, which is a package used for scientific computing. It allows us to define constraints and objectives for the MILP model through indirect linear inequations using lists or arrays. Thus, using MILP, the research aims to provide insight on financial data regarding CCS.

1.6 Geographic Information Systems (GIS)

Geographic Information Systems (GIS) are tools used to collect, store, analyze, and display location-based data[24]. GIS primarily works with shapefiles, a vector data format (otherwise uses files of raster format), storing geometric and geographical spatial data[25]. In this research, GIS helps visualize and evaluate different regions based on political features and CCS locations. It allows us to map real-world data and combine it with predictions from ML. GIS has been proven to be extremely useful in mapping CO_2 emission sources and CCS plants to potential sinks[26], which indicates the effectiveness of the tool in this project.

In GIS, layering refers to organizing different geographic data into separate, visual layers[27]. These layers are then stacked onto a base map, which acts as a reference. This method allows users to analyze spatial relationships and patterns by toggling layers on or off and observing how features interact within a particular region. Research by Coulter et al.[28], which evaluated usage of high-resolution imagery for accuracy and efficiency of GIS vegetation layers, found overall accuracies of the GIS vegetation layers to be around 75 to 80 percent with high-resolution digital image data, which underscores the utility of GIS layers in precise visual understanding. In GIS, a general layered architecture includes four layers: the user interface, application logic, interface connectors, and data services layers[29]. This structure is characterized by bidirectional flow of data across layers, data abstraction and independent layers that can be isolated. Spatial interpolation, the estimation of unknown values at specific locations utilizing known data from nearby points, can be performed on the GIS layers. However, spatial interpolation has not been used in this study.

Bubble maps, a type of GIS maps, use circles of varying sizes, colors and transparencies placed on geographic coordinates to represent data values across regions. They are visually intuitive and effective for comparing data, though large value ranges can cause issues in overlapping of bubbles. Bubble maps have been found to effectively visualize spatial and spatio-temporal data, enabling easy interpretation of patterns[30]. This project has utilized bubble maps for effective visualization and global comparison with respect to cost and output power. This approach offers insights into optimal CCS site locations based on potential investment, output capacity, and a balance of both the variables. The research has also superimposed coordinates of pre-existing CCS sites onto a global map obtained as a shapefile from Natural Earth[31], to visually study the dispersion of the sites.

2 Results

2.1 Data

Given the high dimensionality of the input data, with 503 columns across quantitative variables such as latitude, longitude and cost, and categorical features such as country location, currency name, and scope name, the training efficiency of the neural network is a criterion of concern.

2.2 Model

While building the ML model, a multi-layer artificial neural network (ANN) was developed using Python and TensorFlow's Sequential API. The Adam optimizer was utilized, with a learning rate of 0.001 and a batch size of 32. Among several trial configurations, the architecture consisting of an implicit input layer, six hidden layers of decreasing neuron counts – 250, 200, 150, 100, 50 and 25, and finally an output layer with a single node, produced the best and most consistent results. The hidden layers employ Rectified Linear Unit (ReLU) which allows the model to learn non-linear relationships between variables.

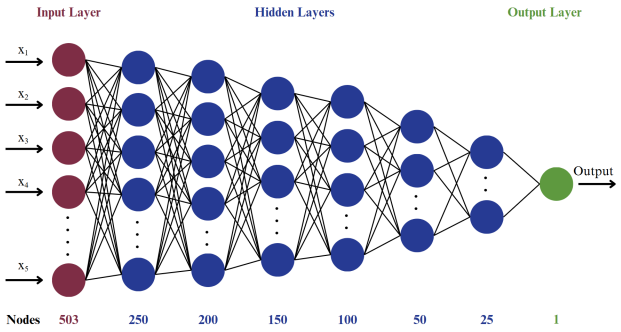

Fig. 3 Architecture of the most Robust ML Model Built

Among the various epoch counts used to train the model, the lowest Mean Squared Error (MSE) was observed to be 626.2424 at 50 epochs, when trained with 50 epochs. On different test cases, the model exhibited slight overfitting, reaching similar minimum MSE values — such as 628.77 —around the 45th epoch out of 50. Despite the mild signs of overfitting, this configuration outperformed other constructs, indicating strong predictive potential with some regularization to improve the generalization. Thus, the reported values of power in megawatts from existing sites in the dataset was utilized to make comparable predictions based on site location information. The dataset employed includes data points ranging from 0-10000 MW, with a mean of 608 MW and standard deviation of 913 MW. With MSE values around 626, the resulting root mean squared error (RMSE) of approximately 25.02 is highly favorable, as it suggests that predictions for capture amount per unit base power deviate by about $\pm$25 megawatts — a relatively small error considering the wide range of the variables.

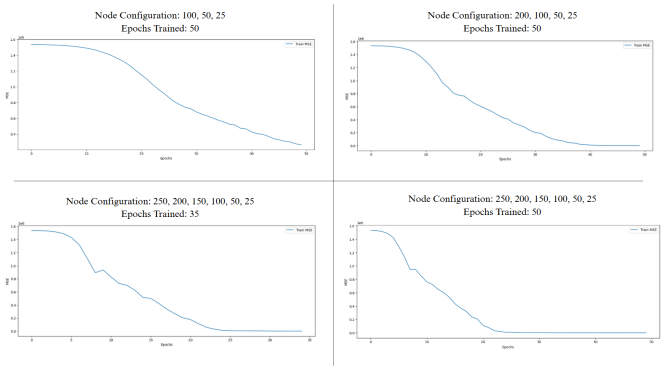

Fig. 4 Comparative MSE Trends across Epochs during Model Training

To reduce overfitting in the neural network, Least Squares Regularization, abbreviated as L2 regularization, was applied, which adds a penalty to large weights during training to encourage simpler and more generalizable models. L2 regularization works by modifying the loss function, helping prevent the model from fitting outliers in the data, which typically results in slightly higher training error (MSE) but better performance on unseen data. After experimentation on the built model, it was found that a regularization strength of 0.000095 yielded the best balance, with an MSE of 829.45 at epoch 48 – statistically less overfitting compared to the unregularized model. Figure 5 depicts the reduction in overfitting with L2 regularization. Alternative regularization techniques such as dropout and early stopping were not utilized in this research.

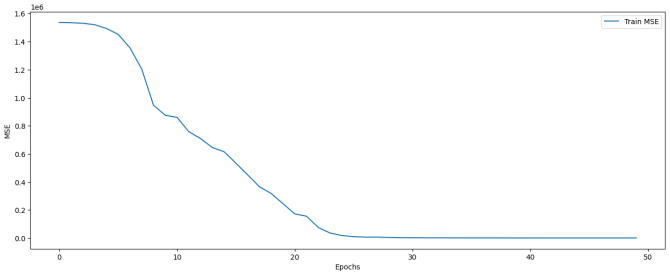

Fig. 5 MSE vs Epochs Trained for Model with Seven Dense Layers and L2 Regularization

In addition to the MSE values, the MAE for the model's ideal scenario with the 7-layer configuration decreasing from 250 nodes to 1 output node trained over 50 epochs was found to be 6.62, which indicates a significantly low error in the predictions given the wide range of data. The R^2 value for the model was 0.99858, implying that the model explains over 99.85% of the variation in the data. These results demonstrate the model's high accuracy and consistency across all the evaluated metrics.

Moreover, SVMs and decision trees were utilized to compare their predictive accuracy with ANNs. The SVM with a linear kernel achieved an RMSE of 968.28 MW, the lowest among the three kernels implemented, whereas the polynomial kernel had an RMSE of 1066.47 MW and the radial basis function (RBF) kernel had an RMSE of 1066.51 MW. The decision tree demonstrated a stronger performance, with an RMSE of 387.71 MW. Figure 6 depicts the structure of the decision tree used in this study. However, these results still underperform compared to the accuracy achieved by ANNs, highlighting their effectiveness in predicting CCS site viability.

The large dimensionality of the input layer is primarily attributed to the number of indicator variables utilized to facilitate regression with categorical variables. According to a study by Manry et al.[32], backpropagation in epochs – the main method used to train neural networks – grows super-linearly with the number of inputs, leading to significant computational costs. To make training faster, the researchers have suggested using data compression or dimensionality reduction techniques, such as linear transformations to reduce the input complexity. Furthermore, the research demonstrated that neural networks, when appropriately sized and well-trained, can approximate the minimum

Table 1 Training Configuration and Performance Metrics of the Neural Network

Dense Layer Count	Node Configuration	Epochs Trained	MSE (MW^2)	Epoch of Optimal MSE	Model Training Time (seconds)
4	100, 50, 25, 1	50	182675.3906	49	9
5	200, 100, 50, 25, 1	50	1023.1886	50	10
6	200, 150, 100, 50, 25, 1	50	804.5150	49	12
7	250, 200, 120, 100, 50, 25, 1	50	680.6991	50	13
7	250, 200, 150, 100, 50, 25, 1	35	1314.7893	35	10
7	250, 200, 150, 100, 50, 25, 1	45	927.2125	40	13
7	250, 200, 150, 100, 50, 25, 1	50	626.2424	50	14
7	250, 200, 150, 100, 50, 25, 1 (With L2 regularization, strength = 0.000095)	50	829.4526	48	16

Table 2 RMSE Values of Machine Learning Algorithms Employed

ML Algorithm	RMSE (MW)
ANN	25.02
SVM	
Linear Kernel	968.28
Polynomial Kernel	1066.47
Radial Basis Function Kernel	1066.51
Decision Tree	387.71

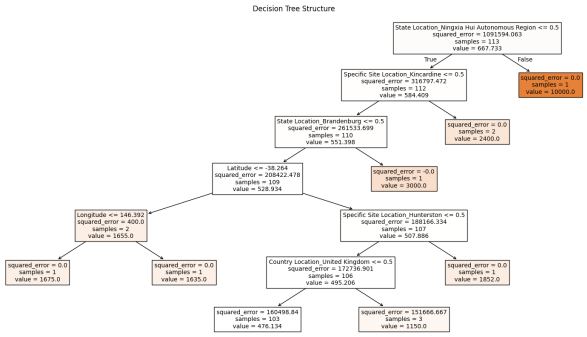

Fig. 6 Decision Tree Created

mean square error with high accuracy. This suggests that neural networks are well-suited for this task though no dimensionality reduction techniques were used in this research since future work with the model built could involve improving performance by compressing input features prior to training.

In order to understand the relationships among the variables in the input layer, correlation analysis was conducted on the quantitative and qualitative features. A heatmap visualization was used to assess the quantitative variables, which indicated a very low correlation between longitude and latitude values, as seen in Figure 7. For the categorical variables, the chi-squared test was applied to examine the relationships among country location, state location, specific site location, and continent name. The results, as depicted in Figure 8, illustrated that specific site location and continent name were independent, with a p-value of approximately 0.185, which exceeds the standard threshold p-value of 0.05. In this case, the null hypothesis was accepted, that is, there was no correlation between the variables. Contrastingly, all other pair-wise combinations yielded p-values significantly below 0.05, indicating relationships between them. This outcome aligns with the expectations since all the combinations were related to the geographic locations of sites.

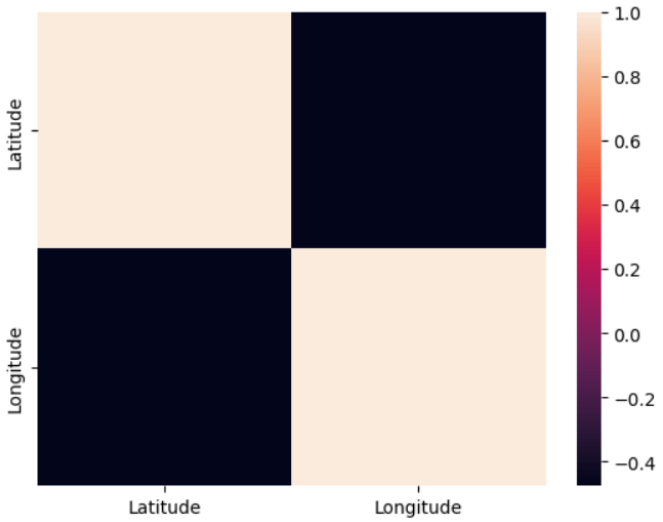

Fig. 7 Heatmap between Latitude and Longitude

The SHAP values of the variables were calculated to understand each feature's impact on the prediction. Due to the large number of input features, the SHAP values of all the 503 variables were not plotted. Figures 9 and 10 depict the SHAP values plotted in a bar graph and a beeswarm graph respectively. The values for 9 variables are explicitly plotted, along with the sum of the remaining 494 features.

Variable 2 Variable 1	Continent Name	Country Location	Specific Site Location	State Location
Continent Name	NaN	5.565665e-64	1.850902e-01	3.708709e-14
Country Location	5.565665e-64	NaN	7.579510e-03	5.209348e-95
Specific Site Location	1.850902e-01	7.579510e-03	NaN	7.217933e-07
State Location	3.708709e-14	5.209348e-95	7.217933e-07	NaN

Fig. 8 Chi-Squared Test Results between Categorical Variables

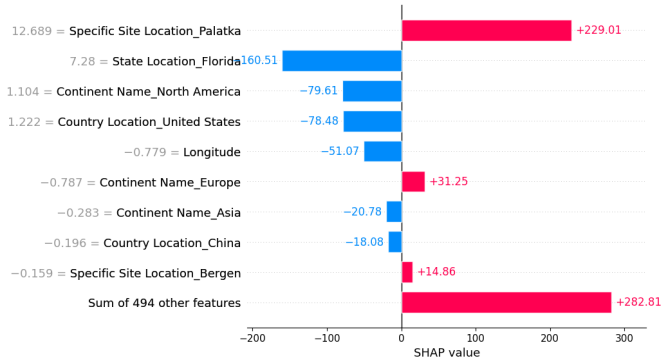

Fig. 9 Bar Plot of SHAP Values

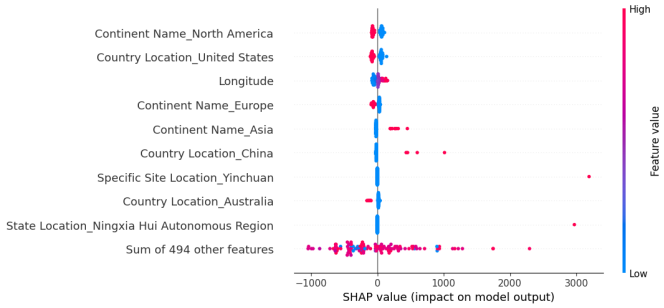

Fig. 10 Beeswarm Plot of SHAP Values

Mixed Integer Linear Programming (MILP) was attempted using CCS cost and capture amount per unit base power as variables within linear inequality constraints. However, the problem was found to be unbounded – while the simplex algorithm identified a feasible solution, it also found a direction where all variables could increase while still reducing the objective cost, which made the problem unsolvable in its current form. This indicates that the MILP model was not viable in this research but could prove useful in future works that included specific constraints and data inputs.

As a visual alternative to MILP, the research utilized GIS alongside cost and power data. A global map was created to display existing CCS sites, offering a visual reference of regions where suitable infrastructure and transport networks already exist. One key observation from this visualization was that the majority of present CCS sites were located along coastlines. This tendency is supported by a study on CCS deployment in Finland, which concluded that the most viable sites are those located on the coast and with large volumes of CO_2 [33]. A coastal location allows for efficient CO_2 transport by ship, making it a highly practical option during the early phases of CCS implementation. This insight is likely applicable on a global scale.

Another study on the restoration of coastal CCS plants, specifically using seagrasses, found that even modest rates of recovery could boost global CCS capacity by nearly 10% over the next century [34]. These studies, when combined with the map-based data, help inform future site selection by highlighting practicality of coastal locations from an industrial perspective.

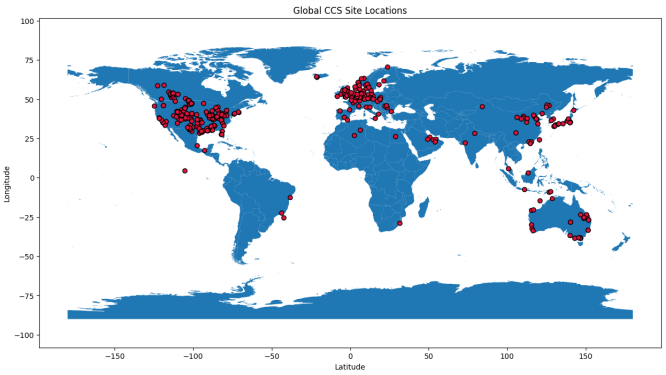

Fig. 11 Global Distribution of Existing CCS Sites

Additionally, global bubble maps were created to visualize and compare costs (in millions), power production (in megawatts), and power efficiency (megawatts produced per million spent). These maps not only help identify cost-effective and high-output regions but also allow for spatial comparisons between countries or continents. They can be used to prioritize locations for future investment or research and guide decisions based on geographical and economic suitability. Figure 14 indicates the power efficiency or the power produced in megawatts per million US Dollars spent. This was calculated by standardizing the cost data in the dataset in terms of 1 million USD and then obtaining the ratio between the size or capture amount per unit MW to the cost in millions (USD).

Overall, the results show consistent model performance and valuable geographic patterns in CCS site distribution, offering both predictive accuracy and site viability. This addresses the research question by highlighting the effectiveness of ML models in CCS site prediction, when integrated with tools such as GIS.

3 Discussion

The study moved beyond traditional, location-specific assessments and toward a more scalable and predictive framework that supports global CCS deployment. By drawing from publicly available datasets, this research built a predictive model

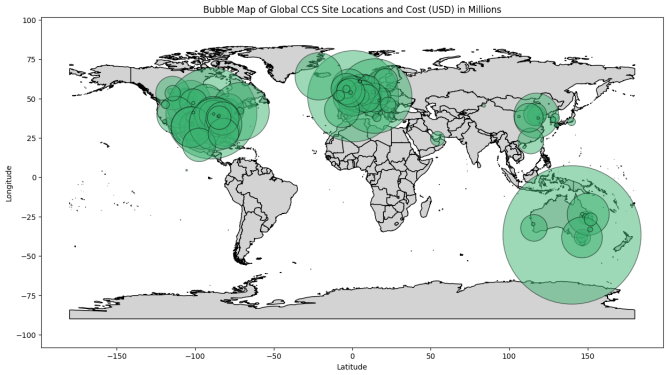

Fig. 12 Global Distribution of CCS Site Costs (in Millions USD)

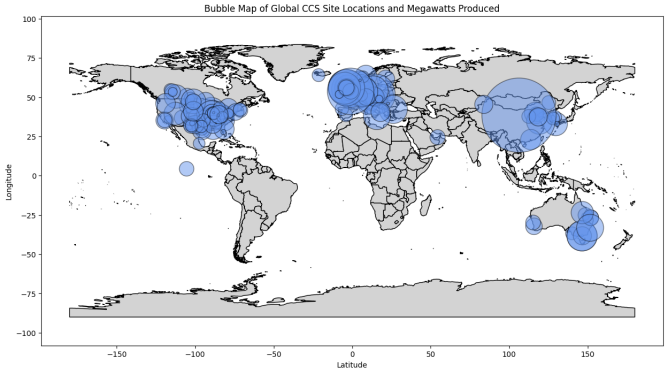

Fig. 13 Global Distribution of Power Output at CCS Sites (in Megawatts)

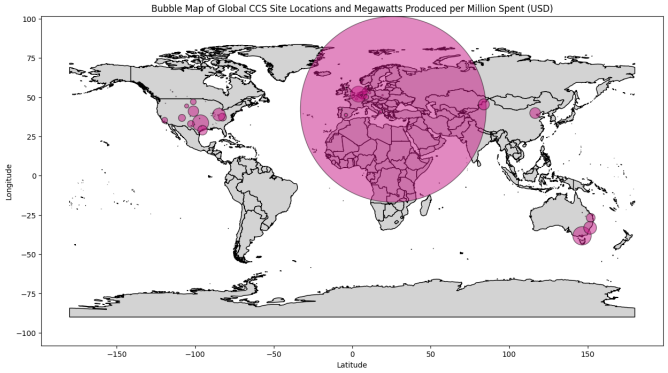

Fig. 14 Global Distribution of CCS Site Power Efficiency (Megawatts per Million USD)

that identified patterns and optimize site selection. One of the central goals was to assess whether ML can deliver accurate predictions of site suitability when multiple variables are considered in tandem. Integrating certain tools such as Mixed Integer Linear Programming (MILP) along with Geographic Information Systems (GIS) into the model would have enhanced the robustness of results. Conversely, removing these layers would have reduced the overall accuracy, reinforcing the importance of a multidimensional approach to decision-making. While the model does not incorporate every possible risk factor, such as potential CO_2 leakage or long-term monitoring costs, it is designed to deliver generalizable insights on site viability. Moreover, the model in this study provided predictions for capture amount per unit base power given that it was the only quantitative metric in the dataset that indicated site suitability. This reflects a limitation in the scope of publicly available data and highlights the need for more comprehensive reporting on CCS sites. However, variables such as storage capacity and injectivity can be utilized in future work.

This work builds on the foundation laid by prior studies that primarily emphasized geological modeling. However, this research distinguishes itself by treating variables in isolation, where the model treats each existing site as a data point, considering how overlapping factors influence CCS potential using visual data. In doing so, it reflects real-world conditions and acts as a tool that may eventually support both policymakers and industries in identifying high-priority storage sites.

Despite its strengths, the project presents its own set of challenges. The process of collecting and analyzing large-scale datasets presents risks such as inconsistencies, missing values, and varying levels of data availability across regions. These inconsistencies may have affected the reliability of predictions. Moreover, integrating different computational tools such as GIS platforms, ML algorithms, and MILP required careful data preprocessing to ensure compatibility. To address these concerns, specific methods were utilized. For example, using multiple datasets (training and testing sets) allowed validation, while preprocessing steps like categorizing and standardizing variables improved overall efficiency and performance. Starting with simpler models and scaling up gradually also helped manage complexity effectively.

A study by van den Broek et al.[8] utilized MILP and sourced its cost data from another research by Damen et al.[35], which focused on CCS deployment pathways, estimating storage capacity, emissions reduction potential, and associated economic requirements. In their 2010 study, van den Broek et al. used energy efficiency and total capital requirement for cost inputs. In contrast, the dataset in this project included overall project costs from various regions, often in different currencies. This made it complex to define clear, standardized constraints for the algorithm and as a result, the MILP setup could not be applied effectively in this research.

In the dataset used in this study, 197 entries detailed capture amount per unit base power. However, a unit inconsistency was observed: while 173 of the records listed power output in megawatts (MW), the remaining entries used alternative units that were incompatible for direct conversion. These units often referred to fundamentally different dimensions, making standardization impractical without introducing errors. As a result, the 24 incompatible records were filtered out to maintain consis-

tency during preprocessing and although necessary, this filtering led to a reduction in the volume of usable data for model training and testing. However, the data points lost had a uniform distribution globally (for example, 2 data points each were lost from USA and UAE), thus not impacting the model significantly.

Moreover, the dataset utilized is unbalanced and therefore, the model is biased towards North America and Europe. This disproportion is evident from the SHAP values which indicate that the categorical features pertaining to North America and Europe significantly impacted the result. This issue, however, is not limited to the dataset used in this study, and has been a concern in prior research. A study by Wang et al.[36] on a tool to mitigate biases in visual datasets found over-representation of data from North America and Europe as compared to other regions. Another study by Beck et al.[37] acknowledged that biases in environmental datasets can arise from differences in the extent data governments and their agencies collect and report. Moreover, the cost-related data utilized in the study were not adjusted to inflation.

Although this study focuses on the effectiveness of ML in CCS site selection, other aspects such as geo-political implications and environmental concerns must be considered. Equity in CCS site selection can be compromised by biases in the dataset, which can be deterred through effective communication between governments and surveyors to ensure collection of accurate data. Environmental risks, such as groundwater contamination from leakage, air pollution, and increased emissions from transport of CO_2, can be addressed by deploying advanced monitoring technologies to detect and prevent leakage, and utilizing green transport.

As the research progressed, one of the key outcomes encountered was the development of a model that demonstrated strong predictive performance, as evaluated by low Mean Squared Error (MSE) values. Even if some variables remain unaddressed in the current scope, such as leakage risks or long-term monitoring, the approach offered valuable general insights into site suitability. Moreover, the methodology employed serves as a framework for future research that aim to expand on this work by incorporating additional environmental or technical factors.

Looking ahead, there are several promising directions this research could take. One application involves using the model's predictions to identify areas where CCS-supportive policies and infrastructure investments would be most impactful. Moreover, industries located near the predicted high-potential CCS regions could collaborate with governments or climate organizations to implement CO_2 emission reduction strategies.

In essence, the research highlights the potential of machine learning in CCS as well as the importance of integrating different factors to enable practical site selection. The power of ML in pinpointing CCS sites is undeniable. However, its true value lies not just in prediction, but in empowering us to make informed choices that secure a sustainable future, one carefully chosen site at a time.

In conclusion, this study demonstrates the effective application of a multi-layer artificial neural network combined with GIS visualization to predict and analyze optimal carbon capture and storage (CCS) site suitability on a global scale. Despite challenges such as data inconsistencies, overfitting, and the limitations encountered with MILP modeling, the developed ANN model achieved promising predictive accuracy with relatively low MSE values and margins of error as low as 25 megawatts, highlighting the potential of machine learning to predict complex, nonlinear relationships among geological and economic variables. The integration of GIS spatial data further deepened the analysis by revealing practical geographic patterns, particularly the prominence of coastal sites, which aligns with existing research. While the current model does not encompass all risk factors yet, it provides a robust, scalable framework that can aid future CCS site selection, thus contributing to the strategic advancement of carbon mitigation efforts worldwide.

Acknowledgment

I would like to express my sincere gratitude to my mentor, Emily Sheetz, for her constant guidance and invaluable insights throughout this project. I also thank the Lumiere Research Program and its staff for their support and resources. Finally, I am deeply grateful to my family for their encouragement and unwavering support during the writing of this paper.

References

1 M. Salam and T. Noguchi, *Impact of human activities on carbon dioxide (CO2) emissions: A statistical analysis*.

2 I.E.A., *Global energy review 2025: CO2 emissions*, https://www.iea.org/reports/global-energy-review-2025/co2-emissions.

3 N. Grid, *What is carbon capture and storage?*, https://www.nationalgrid.com/stories/energy-explained/what-is-ccs-how-does-it-work.

4 *Swiss Federal Office for the Environment (FOEN*, https://www.bafu.admin.ch/bafu/en/home/topics/climate/info-specialists/emission-reduction/reduction-targets/2050-target.html, 2025, January 3). 2050 net-zero target.

5 *United Nations Framework Convention on Climate Change*, https://unfccc.int/process-and-meetings/the-paris-agreement.

6 *Climate Action Tracker*, https://climateactiontracker.org/global/emissions-pathways/.

7 W. Ashraf and V. Dua, *Machine learning based modelling and optimization of post-combustion carbon capture process using MEA supporting carbon neutrality*, https://doi.org/10.1016/j.dche.2023.100115.

8 M. Broek, E. Brederode, A. Ramírez, L. Kramers, M. Kuip, T. Wildenborg, W. Turkenburg and A. Faaij, *Designing a cost-effective CO2 storage infrastructure using a GIS based linear optimization energy model*, https://doi.org/10.1016/j.envsoft.2010.06.015.

9 K. Banachewicz, *Carbon capture and storage, Active, proposed, and terminated CCS projects worldwide. Kaggle*, https://www.kaggle.com/datasets/konradb/carbon-capture-and-storage/data.

10 I.B.M., *What is machine learning?*, https://www.ibm.com/think/topics/machine-learning.

11 I.B.M., https://www.ibm.com/think/topics/reinforcement-learning, What is reinforcement learning?

12 I.B.M., *What is a neural network?*, https://www.ibm.com/think/topics/neural-networks.

13 GeeksforGeeks, *Decision Tree*, https://www.geeksforgeeks.org/machine-learning/decision-tree/.

14 Y. Yan, T. Borhani, S. Subraveti, K. Pai, V. Prasad, A. Rajendran, P. Nkulikiyinka, J. Asibor, Z. Zhang, D. Shao, L. Wang, W. Zhang, Y. Yan, W. Ampomah, J. You, M. Wang, E. Anthony, V. Manovic and P. Clough, *Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS) – a state-of-the-art review*, https://doi.org/10.1039/D1EE02395K.

15 B.-D. Helmy, *The C Parameter in Support Vector Machines*, https://www.baeldung.com/cs/ml-svm-c-parameter.

16 GeeksforGeeks, *Gamma Parameter in SVM*, https://www.geeksforgeeks.org/gamma-parameter-in-svm/.

17 S. Afaq and S. Rao, *Significance of epochs on training a neural network*.

18 M.-C. Popescu, V. Balas, L. Perescu-Popescu and N. Mastorakis, *Multilayer perceptron and neural networks*.

19 K. Hornik, M. Stinchcombe and H. White, *Multilayer feedforward networks are universal approximators*, https://doi.org/10.1016/0893-6080(89)90020-8.

20 N. Sipöcz, F. Tobiesen and M. Assadi, *The use of artificial neural network models for CO2 capture plants*, https://doi.org/10.1016/j.apenergy.2011.01.013.

21 G. Cybenko, *Approximation by superpositions of a sigmoidal function*, https://doi.org/10.1007/BF02551274.

22 M. Stojiljković, *Hands-on linear programming: Optimization with Python*, https://realpython.com/linear-programming-python/#linear-programming-explanation.

23 F. Zhang, E. Catalanotti, S. Martynov, R. Porter and H. Mahgerefteh, *A mixed-integer linear programming model for multi-modal CO2 transport*, https://doi.org/10.2139/ssrn.5059008.

24 N. Geographic, https://education.nationalgeographic.org/resource/geographic-information-system-gis/, GIS (Geographic Information System).

25 U. Delaware, *File Formats for GIS*, https://sites.udel.edu/gis/file-formats-for-gis/.

26 A. Yousefi-Sahzabi, K. Sasaki, I. Djamaluddin, H. Yousefi and Y. Sugai, *GIS modeling of CO2 emission sources and storage possibilities*, https://doi.org/10.1016/j.egypro.2011.02.188.

27 MangoMap, *GIS Mapping*, https://mangomap.com/gis-mapping.

28 L. Coulter, D. Stow, A. Hope, J. O'Leary, D. Turner, P. Longmire, S. Peterson and J. Kaiser, *Comparison of high spatial resolution imagery for efficient generation of GIS vegetation layers*.

29 X. Bian, R. Hu, Z. Yu and D. Li, *Study on layered architecture model for distributed GIS*.

30 M. Shaito and R. Elmasri, *Map visualization using spatial and spatio-temporal data: Application to COVID-19 data*, https://doi.org/10.1145/3453892.3461336, ACM.

31 N. Earth, *Admin 0 – Countries*, https://www.naturalearthdata.com/downloads/50m-cultural-vectors/50m-admin-0-countries-2/.

32 M. Manry, S. Apollo and Q. Yu, *Minimum mean square estimation and neural networks*, https://doi.org/10.1016/0925-2312(95)00101-8.

33 S. Teir, E. Tsupari, A. Arasto, T. Koljonen, J. Kärki, A. Lehtilä, L. Kujanpää, S. Aatos and M. Nieminen, *Prospects for application of CCS in Finland*, https://doi.org/10.1016/j.egypro.2011.02.628.

34 A. Irving, S. Connell and B. Russell, *Restoring coastal plants to improve global carbon storage: Reaping what we sow*, https://doi.org/10.1371/journal.pone.0018311.

35 K. Damen, A. Faaij and W. Turkenburg, *Pathways towards large-scale implementation of CO2 capture and storage: A case study for the Netherlands*, https://doi.org/10.1016/j.ijggc.2008.07.002.

36 A. Wang, A. Narayanan and O. Russakovsky, *REVISE: A tool for measuring and mitigating bias in visual datasets*, https://doi.org/10.1007/978-3-030-58580-8_43.

37 L. Beck, T. Bernauer and A. Kalbhenn, *Environmental, Political, and Economic Determinants of Water Quality Monitoring in Europe*.