

# Predictive Modeling of CO<sub>2</sub> Wettability at Brine-Mineral Interfaces as a Mechanistic Framework for Sustainable Carbon Storage

Mohammadreza Kazemi

kazemi.bipc@gmail.com

Petrochemical Research and Technology Company (Iran)

---

## Article

**Keywords:** Carbon dioxide sequestration, Interfacial wettability, Machine learning, CatBoost algorithm, SHAP interpretability

**Posted Date:** September 21st, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10932488/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Predictive Modeling of CO<sub>2</sub> Wettability at Brine-Mineral Interfaces as a Mechanistic Framework for Sustainable Carbon Storage

Mohammadreza Kazemi<sup>2,\*</sup>

<sup>1</sup>Process Engineering Department, Bandar Imam Petrochemical Company (BIPC), Mahshahr, Iran.

\*Corresponding author: kazemi.bipc@gmail.com

## Abstract

Accurately predicting the interfacial wettability of carbon dioxide in deep saline aquifers is critical for ensuring the secure structural and residual trapping of sequestered carbon. This study aims to develop and systematically evaluate advanced machine learning algorithms to reliably forecast the equilibrium contact angle at the carbon dioxide-brine-mineral interface under diverse thermodynamic and geochemical conditions. A comprehensive dataset comprising 630 empirical observations of multiphase interfacial behaviors across varied mineralogical compositions, including quartz, calcite, feldspar, and mica, was curated. Fifteen diverse predictive models, ranging from fundamental linear regressions to advanced deep learning and tree-based ensembles, were trained and validated utilizing a rigorous five-fold cross-validation and grid search hyperparameter optimization framework. Performance evaluations revealed that traditional linear architectures underfit the complex non-linear thermodynamic interactions, whereas basic decision trees exhibited severe overfitting. Gradient boosting frameworks demonstrated superior generalization, with the CatBoost algorithm emerging as the optimal model, achieving an exceptional testing  $R^2$  of 0.9357 and a minimized MSE of 20.1358. Advanced visual diagnostics, including Taylor diagrams, parity dispersion plots, and residual probability density functions, confirmed the spatial concordance and minimized systematic bias of the CatBoost architecture. Furthermore, SHAP interpretability analysis identified operational pressure, baseline wetting state ( $\theta_0$ ), and mica weight fraction as the paramount physical parameters governing the predictive output. The deployed machine learning methodology successfully provides a robust and mathematically stable predictive framework for modeling complex interfacial wettability. Integrating these advanced computational algorithms optimizes site-selection criteria and enables the dynamic calibration of injection strategies, thereby fundamentally enhancing macroscopic storage security in deep geological carbon sequestration operations.

**Keywords:** Carbon dioxide sequestration, Interfacial wettability, Machine learning, CatBoost algorithm, SHAP interpretability

## 1. Introduction

The main cause of climate change in recent years has been the release of CO<sub>2</sub> into the atmosphere. Driven by industrial progress and rising energy demands, these emissions are expected to continue growing [1]. With the aim of curbing CO<sub>2</sub> emissions, regulatory frameworks have been developed. Toward this objective, diverse methods, including geological CO<sub>2</sub> sequestration and the use of biofuels in replacement of fossil fuels were recommended [2]. The capture and geological storage of CO<sub>2</sub> is recognized the most highly recommended method for reducing the atmospheric greenhouse effect [3-5].

Through a detailed numerical simulation, Chen et al. examined improved water recovery by CO<sub>2</sub> injection (CO<sub>2</sub>-EWR) alongside carbon storage and capture technologies. Their research focused on assessing various CO<sub>2</sub> injection scenarios to maximize both CO<sub>2</sub> storage and brine production. They noted that combining co-production and preproduction of water in CO<sub>2</sub>-EWR systems offers significant advantages, including improved CO<sub>2</sub> storage and better control of pressure perturbations. Despite this, the cost of drilling additional wells to increase CO<sub>2</sub> storage capacity was found to surpass the economic gains from implementing the method [6]. Jing et al. explored how salinity and dip angle affect the storage mechanisms and volume of CO<sub>2</sub> in geological storage systems. Using a 3D simulation model, they determined that higher salinity correlates with decreased storage capacity, and greater dip angles contribute to longer CO<sub>2</sub> migration distances. They recommended that reservoirs with lower salinity and smaller dip angles are preferable for optimal CO<sub>2</sub> storage [7].

Various mechanisms have been introduced in the literature to inhibit CO<sub>2</sub> leakage, with structural trapping [8], residual [9], solubility [10], and mineral trapping [11]. The effectiveness of these mechanisms varies depending on the characteristics of the aquifer and cap rock. While each

mechanism is active at different time scales, structural and residual trapping are regarded as the primary mechanisms influencing the early phases of CO<sub>2</sub> storage [12, 13]. Structural trapping involves the containment of CO<sub>2</sub> underneath a low-permeability seal or cap rock. Residual trapping, on the other hand, occurs in formations without a cap rock and depends on capillary forces to stop CO<sub>2</sub> mobilization. The effectiveness of this mechanism is highly sensitive to initial CO<sub>2</sub> saturation, rock morphology, rock wettability, and CO<sub>2</sub>/brine interfacial tension. As a result, gaining detailed insights into wettability is essential for assessing residual trapping potential and forecasting CO<sub>2</sub> leakage risks [14-16].

Wettability is a key factor in petroleum engineering, and its behavior is predominantly shaped by environmental conditions such as salinity, pressure, and temperature. Various studies have explored the contact angles of CO<sub>2</sub>/brine on different mineral surfaces. Farokhpour et al. specifically specified the contact angles of CO<sub>2</sub>/brine on mica, feldspar, quartz, and calcite under varying conditions. Their findings demonstrated that pressure reduced the water-wetness of mica, shifting it toward weaker water-wet conditions, whereas other mineral elements showed little variation. Moreover, they observed that the maximum contact angle for feldspar, quartz, and calcite at critical pressure was approximately 36° [17]. Chen et al. conducted molecular dynamics simulations to examine the water contact angle on silica surfaces under different pressure (2.8–32.6 MPa) and temperature (318–383 K) conditions. The study demonstrated that the effects of temperature and pressure on the contact angle are influenced and potentially controlled by the characteristics of surface functional groups [18]. Chen investigated the receding, static, and advancing contact angles of CO<sub>2</sub>/brine on quartz under varying conditions of salinity, pressure, and temperature. Molecular dynamics simulations were then employed to model the wettability behavior. According to their findings, it was concluded that temperature and pressure have minimal

impact on contact angles, whereas ionic strength significantly surges water contact angle. The modeling results demonstrated excellent arrangement with the experimental data [19]. Mutailipu established an empirical model to predict the CO<sub>2</sub>-brine interfacial tension (IFT) in relation to pressure, salinity, and temperature. Moreover, they studied contact angle values on Berea sandstone, quartz, and limestone for liquid, gaseous, and CO<sub>2</sub>. It was observed that the wettability of limestone and Berea sandstone transitioned to smaller water-wet conditions, whereas quartz demonstrated a dissimilar response under supercritical circumstances [20].

Accurately characterizing the interfacial wettability of carbon dioxide within deep saline aquifers is fundamentally critical for ensuring the long-term containment security and volumetric efficiency of geological carbon sequestration. Without a precise quantitative understanding of these multiphase dynamics, assessing the viability of structural and residual trapping mechanisms remains highly uncertain, potentially compromising large-scale climate mitigation operations. The core problem addressed in this study is that the inherent complexity and highly non-linear interactions among varying thermodynamic states, variable brine salinities, and heterogeneous mineralogical compositions render traditional mathematical or strictly empirical estimations of the equilibrium contact angle inadequate and prohibitively resource-intensive. Consequently, this research tests the hypothesis that advanced, data-driven machine learning frameworks can overcome these structural limitations by deciphering high-dimensional physical interactions to predict wettability angles with significantly higher fidelity than conventional predictive architectures. To test this hypothesis and establish a robust modeling paradigm, the research workflow begins with aggregating and preprocessing a comprehensive empirical dataset of thermodynamic and geochemical parameters. The objective is to systematically train, validate, and compare fifteen diverse algorithmic architectures, ranging from fundamental regressions to

advanced ensemble and deep learning networks, using strict hyperparameter optimization and five-fold cross-validation. Finally, the workflow (Figure 1) integrates SHAP-based interpretability analysis to quantitatively isolate the governing physical and mineralogical drivers, thereby providing a fully transparent and field-applicable predictive tool without relying on computationally expensive laboratory replications.

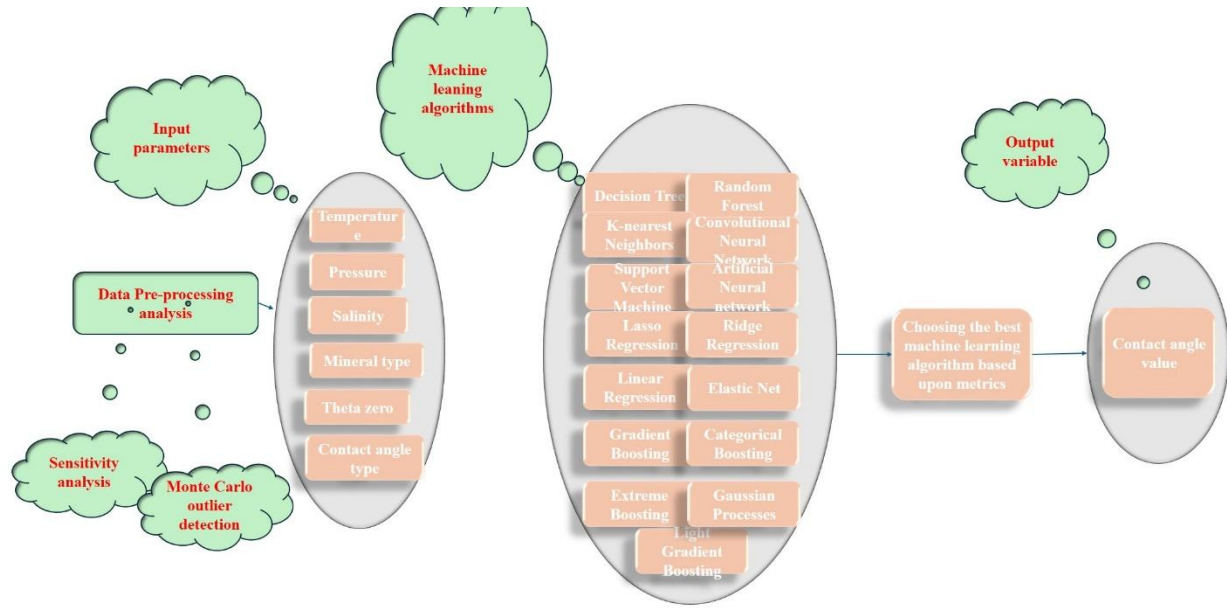


Figure 1. A data-driven approach to model contact angle in this work

## 2.1 Machine Learning Algorithms

Linear Regression (LR) is implemented as a fundamental statistical approach designed to establish a direct linear relationship between the independent variables, such as mineral composition and thermodynamic conditions, and the target contact angle. This technique is widely utilized in preliminary regression analyses to provide a baseline understanding of dataset dynamics [21, 22]. The primary advantage of LR lies in its exceptional computational efficiency and inherent interpretability, allowing for the straightforward extraction of structural feature weights. However, its significant disadvantage is the inability to capture complex, non-linear interactions commonly

observed in multiphase interfacial phenomena. Consequently, LR is selected in this study to serve as a standardized analytical baseline, ensuring that the performance gains achieved by more sophisticated, non-linear algorithms can be systematically quantified and justified against a fundamentally simple mathematical architecture [23, 24].

Ridge Regression is deployed as a regularization-enhanced variant of conventional linear modeling, wherein an L2 penalty term is integrated into the cost function to rigidly constrain the magnitude of the model coefficients. This algorithm is extensively applied in scenarios where datasets exhibit high multicollinearity, a common occurrence when evaluating interdependent variables like advancing and receding contact angles. The salient advantage of Ridge is its capacity to prevent severe overfitting by penalizing inflated weights, thereby enhancing the generalization capability of the model on unseen experimental data. Conversely, a notable disadvantage is that Ridge does not perform absolute feature selection, meaning all input variables are retained in the final algebraic equation regardless of their actual statistical significance [25]. This algorithm is chosen herein to mitigate potential multicollinearity among the diverse geological and thermodynamic inputs while strictly preserving the overall baseline linear framework [26].

Lasso regression is introduced as an advanced regularized regression technique that integrates an L1 penalty, which forces the coefficients of less contributory variables to become exactly zero. This intrinsic mathematical characteristic makes Lasso highly applicable in high-dimensional datasets requiring simultaneous predictive modeling and automated feature selection. The most prominent advantage of Lasso is its inherent ability to yield sparse, highly interpretable models by systematically eliminating irrelevant inputs, thus substantially reducing computational noise. Nevertheless, its primary disadvantage emerges when dealing with highly correlated independent variables, as Lasso tends to arbitrarily select one feature while immediately discarding the others,

potentially omitting valuable synergistic information. Lasso is strategically selected for this comprehensive investigation to identify and isolate the most dominant thermodynamic and mineralogical parameters governing CO<sub>2</sub> wettability, thereby streamlining the predictive architecture while maintaining strict analytical rigor [26, 27].

Elastic Net (EN) is formulated as a hybrid regularization technique that mathematically combines the L1 and L2 penalty terms utilized in Lasso and Ridge, respectively. This integrated approach is particularly applicable in complex regression tasks where both model sparsity and robustness against extreme multicollinearity are simultaneously demanded. The paramount advantage of EN is its proven ability to overcome the stringent limitations of Lasso by grouping highly correlated variables together, ensuring that synergistic statistical effects among inputs like pressure and salinity are preserved rather than arbitrarily discarded. A limiting disadvantage is the introduction of additional hyperparameters that must be meticulously tuned, substantially increasing the computational overhead during the model training phase. EN is specifically incorporated into this analytical framework to harness a balanced, dual-regularization strategy, ensuring optimal utilization of the interdependent input features dictating solid-fluid interfacial behavior [28, 29].

Support Vector Regression (SVR) is implemented as a robust computational algorithm that operates by projecting input features into a highly elevated dimensional space using specialized kernel functions to construct an optimal regression hyperplane [30, 31]. This sophisticated methodology is extensively applied in complex engineering and chemical modeling where the underlying relationships are highly non-linear and multidimensional. The principal advantage of SVR is its exceptional resilience to statistical outliers and its capability to define a precise mathematical margin of tolerance, thereby ensuring a highly generalized predictive performance. However, a significant disadvantage is its extreme sensitivity to hyperparameter configuration and

kernel selection, which can drastically escalate the computational cost for massive geological datasets. SVR is purposefully chosen for this wettability analysis due to its proven efficacy in handling complex, non-linear thermodynamic interactions at the brine/CO<sub>2</sub>/mineral interface, providing a highly reliable predictive topology [32, 33].

Random Forest (RF) is constructed as a sophisticated ensemble learning methodology that aggregates the numerical predictions of numerous independent decision trees generated via bootstrap aggregating mechanisms. This versatile computational algorithm is universally applied across diverse scientific disciplines for both classification and complex regression tasks due to its robust statistical foundation [34-36]. The core advantage of RF is its inherent resistance to structural overfitting and its exceptional capability to manage complex datasets without requiring exhaustive preliminary data preprocessing or strict numerical scaling. Conversely, the primary disadvantage of RF is its structural opacity, acting largely as a black-box analytical model that complicates the direct extraction of explicit algebraic relationships between independent inputs and outputs. RF is integrated into this comparative study because of its superior predictive stability and its capacity to efficiently model the highly non-linear dependencies among mineralogical fractions and operating conditions without structural bias.

Gradient Boosting (GB) is developed as an advanced, sequential ensemble technique where subsequent weak learners are systematically trained to minimize the residual predictive errors generated by preceding models. This iterative mathematical approach is widely applied in predictive analytics where achieving the highest possible accuracy is strictly prioritized over computational simplicity [37, 38]. The foremost advantage of GB lies in its profound predictive precision and its ability to flexibly optimize various differential loss functions, making it highly adaptable to specific interfacial dataset characteristics. However, a critical disadvantage is its

profound susceptibility to extreme overfitting if the sequential learning process is not rigorously constrained by appropriate hyperparameters, alongside a relatively slow operational training velocity. GB is carefully selected for this interfacial tension modeling to exploit its aggressive error-reduction architecture, ensuring that subtle, complex variations in contact angles driven by pressure and temperature are meticulously captured.

K-Nearest Neighbor (KNN) is employed as a non-parametric, instance-based learning algorithm that formulates structural predictions by averaging the target values of the closest proximal data points within the multi-dimensional feature space [39, 40]. This fundamentally intuitive algorithm is commonly applied in spatial pattern recognition and baseline regression tasks where the underlying geometric data distribution remains completely unknown. The major advantage of KNN is its structural simplicity and the complete absence of a formal parametric training phase, allowing for immediate deployment and straightforward conceptual interpretation based on Euclidean distances. In contrast, a prominent disadvantage is its extreme mathematical sensitivity to unscaled numerical features and the curse of dimensionality, coupled with high computational latency during the inference phase due to exhaustive distance calculations. KNN is chosen herein to serve as a proximity-based comparative benchmark, rigorously evaluating how well the localized similarities among distinct geological samples can predict unobserved CO<sub>2</sub> wettability characteristics.

Decision Tree (DT) is utilized as a hierarchical, tree-structured analytical algorithm that recursively partitions the multi-dimensional dataset into distinct, non-overlapping regional spaces based on optimized numerical feature thresholds. This methodology is frequently applied in operational research and preliminary data mining due to its highly transparent, rules-based decision-making architecture [41, 42]. The most significant advantage of DT is its unparalleled

structural interpretability, allowing researchers to visually trace the exact logical pathways and mathematical thresholds dictating the final predicted interfacial values. Nevertheless, the predominant disadvantage of DT is its high propensity for severe structural overfitting, as deep, unconstrained trees often memorize statistical noise rather than learning generalized, underlying physical trends. DT is intentionally incorporated into this analytical study to provide explicit, highly interpretable numerical splitting rules that can physically elucidate the hierarchical importance of critical parameters, such as salinity and quartz content, on the final contact angle outcome.

Extreme Gradient Boosting (XGBoost) is deployed as a highly optimized, scalable implementation of the gradient boosting framework, explicitly engineered to maximize computational efficiency and overall model performance. This state-of-the-art computational algorithm is extensively applied in high-stakes predictive modeling and advanced data science due to its robust mathematical handling of sparse and complex datasets. The supreme advantage of XGBoost is its deep integration of advanced L1 and L2 regularization techniques, which systematically mitigate structural overfitting while utilizing automated parallel processing to drastically accelerate the model training phase. However, a notable disadvantage is the sheer mathematical complexity of its hyperparameter space, requiring exhaustive and computationally expensive iterative tuning to achieve an optimal configuration. XGBoost is rigorously selected for predicting CO<sub>2</sub> wettability to aggressively leverage its exceptional computational speed and its unparalleled capacity to extract subtle, highly non-linear thermodynamic patterns from the extensive matrix of mineralogical and experimental data.

Light Gradient Boosting (LightGBM) is introduced as a highly efficient gradient boosting framework that utilizes a unique leaf-wise tree growth strategy rather than the traditional level-

wise approach. This structural algorithmic innovation is widely applied in computational scenarios involving massive, high-dimensional datasets where operational latency and memory consumption are designated as critical constraints. The primary advantage of LightGBM is its unprecedented training velocity and significantly reduced computational memory footprint, successfully achieved without compromising the intrinsic predictive accuracy of the gradient boosting mechanism. Conversely, the main disadvantage of this algorithm is its substantially increased susceptibility to overfitting when applied to relatively small or highly noisy experimental datasets, necessitating stringent parametric control over maximum tree depth and minimum data required per leaf. LightGBM is specifically chosen for this research to facilitate rapid, highly iterative hyperparameter optimization while rigidly maintaining formidable predictive precision regarding the complex physical interactions at the brine/CO<sub>2</sub>/mineral interface.

Categorical Boosting (CatBoost) is implemented as a revolutionary gradient boosting algorithm specifically architected to natively process categorical variables without demanding extensive, preliminary pre-processing techniques like arbitrary one-hot encoding. This powerful methodology is heavily applied in complex datasets encompassing a highly diverse mixture of continuous numerical values and discrete geological classifications. The outstanding advantage of CatBoost is its rigorous utilization of ordered boosting, a sophisticated mathematical mechanism that effectively neutralizes target leakage and significantly enhances the structural generalization capabilities of the developed predictive model. A limiting disadvantage, however, is its considerable computational resource demand during the initial training phase, particularly when processing purely numerical datasets where its specialized categorical optimizations remain entirely unutilized. CatBoost is strategically selected for this comprehensive analytical evaluation to optimally exploit its advanced anti-overfitting mechanisms and structural robustness, ensuring

that the intricate dependencies dictating contact angles are modeled with maximum statistical integrity.

Gaussian Process (GP) is formulated as a sophisticated probabilistic, non-parametric Bayesian approach that rigorously defines a mathematical distribution over possible functions to establish the predictive relationships between complex variables. This mathematically elegant algorithm is primarily applied in spatial statistics and structural optimization tasks where quantifying predictive uncertainty is recognized as being just as critical as the primary prediction itself. The paramount advantage of GP is its inherent ability to provide a rigorous, mathematically sound confidence interval for every generated predicted value, yielding deep structural insights into the operational reliability of the model across different regions of the localized input space. However, a severe disadvantage is its immense computational complexity, which scales cubically with the absolute number of dataset instances, making it highly impractical for extremely large or dynamic geological datasets. GP is deliberately chosen for this wettability framework to probabilistically map the complex interfacial phenomena and provide strict statistical bounds on the forecasted contact angles.

Artificial Neural Network (ANN) is constructed as a highly interconnected computational architecture fundamentally inspired by biological neural networks, comprising multiple distinct layers of artificial neurons continuously optimized through advanced backpropagation algorithms. This foundational deep learning algorithm is universally applied in complex, non-linear regression analysis where the intricate mapping between high-dimensional geological inputs and dynamic outputs defies traditional algebraic formulation. The primary advantage of ANN is its proven universal approximation capability, enabling the continuous mathematical extraction of profoundly abstract and complex representational features from diverse experimental data points.

Conversely, a significant disadvantage is its overwhelming tendency to act as a completely opaque black-box system, compounded by an intense structural vulnerability to overfitting and a heavy operational reliance on vast quantities of clean training data. ANN is critically integrated into this study to maximally harness its immense capacity for modeling the profoundly non-linear, multi-faceted thermodynamic interactions governing CO<sub>2</sub> wettability behavior within distinct mineralogical environments.

Convolutional Neural Network (CNN) is introduced as a specialized deep learning architecture natively designed to process multi-dimensional grid-like topologies, systematically utilizing convolutional mathematical filters to automatically extract hierarchical spatial features [43-45]. While traditionally applied in advanced image recognition tasks, sophisticated one-dimensional variants of this algorithm are increasingly utilized in processing highly structured, sequential, or complex multi-variable analytical regression datasets. The major advantage of CNN is its efficient parameter-sharing mechanism and localized mathematical feature extraction capability, which systematically identify hidden topological correlations among adjacent thermodynamic or compositional input vectors. However, its primary disadvantage lies in the immense computational processing resources required for adequate training and the profound analytical difficulty in physically interpreting the abstracted feature maps generated by the deeper convolutional layers. CNN is selected for this advanced predictive modeling to strictly evaluate its intrinsic capability to detect subtle, sequential numerical patterns within the extensive array of geological inputs dictating interfacial tension phenomena.

## **2.2. Data analysis and explication**

The comprehensive dataset utilized in this analytical framework is systematically extracted from highly regarded, peer-reviewed literature [46, 47] focusing on carbon capture and storage

mechanisms, specifically targeting interfacial wettability phenomena during geological CO<sub>2</sub> sequestration. A substantial portion of the fundamental data points is derived from foundational studies investigating the complex wettability behavior at the brine/CO<sub>2</sub>/mineral interface across diverse mineralogical compositions, including quartz, calcite, feldspar, and mica. The initial baseline behaviors of multiphase systems under varying environmental conditions were critically established by these foundational investigations, whereby the necessity of empirical modeling for resolving the prevailing uncertainties associated with structural and residual trapping capacities in deep geological formations was formally recognized.

Furthermore, supplementary data instances are meticulously integrated from specialized research dedicated specifically to CO<sub>2</sub>/water/quartz wettability behavior within deep saline aquifers, ensuring a highly robust and diverse empirical foundation. By amalgamating these critically evaluated empirical records, a consolidated dataset comprising 630 distinct experimental data points is rigorously constructed, encompassing a wide spectrum of thermodynamic states, fluid salinities, and distinct mineralogical interfaces. This systematic data aggregation is intentionally executed so that the highly nonlinear dynamic variations in interfacial tension are comprehensively captured, thereby providing an optimal, high-fidelity matrix required for training and validating the advanced computational algorithms deployed in this comparative predictive methodology.

The predictive matrix is fundamentally structured upon a highly specific set of independent input variables, prominently featuring operational thermodynamic conditions such as pressure (measured in MPa) and temperature (measured in K), alongside aqueous fluid salinity (measured in M). The inclusion of these specific thermodynamic and chemical parameters is scientifically critical, as the fundamental fluid phase behavior, ionic strength, and electrostatic double-layer interactions occurring at the solid-fluid boundary are directly governed by them. Additionally, the

mineralogical input fractions, specifically quartz, calcite, feldspar, and mica, alongside baseline wettability metrics including static, advancing, and receding indicators, as well as  $\theta_0$  (measured in degrees), are integrated to quantify intrinsic surface energies and initial interfacial states. The accurate mapping of these inputs to the corresponding output is scientifically paramount, as the capillary breakthrough pressures essential for the reliable containment of sequestered CO<sub>2</sub> are inherently dictated by the resulting interfacial behavior.

The designated output parameter for all developed predictive models is the equilibrium contact angle, conventionally quantified in degrees, which is rigorously defined as the geometric angle formulated at the precise intersection where the dense CO<sub>2</sub> phase, the aqueous brine phase, and the solid mineral substrate converge. This specific parameter is widely recognized as the primary macroscopic indicator of the prevailing wettability state, whereby the preferential wetting affinity of the geological formation, whether water-wet, intermediate-wet, or CO<sub>2</sub>-wet, is systematically determined under specific reservoir conditions. The accurate quantification of this output is of immense practical importance for broad industrial applications, as the multiphase flow dynamics, relative permeability, and ultimate efficacy of both residual and structural trapping mechanisms are directly governed by this fundamental property.

To systematically visualize the complex multidimensional interactions and fundamental statistical distributions intrinsic to the aggregated dataset, a comprehensive scatter matrix plot is generated and presented in Figure 2. Within this graphical representation, the univariate kernel density estimations for each respective continuous and categorical variable are explicitly displayed by the principal diagonal elements, highlighting a pronounced positive skewness in operational parameters such as pressure and salinity. Simultaneously, the bivariate scatter distributions between all possible parameter pairs are systematically illustrated by the off-diagonal subplots,

whereby the dense, highly non-linear clustering of the experimental data points is effectively revealed. This comprehensive exploratory data visualization is utilized so that preliminary correlational trends, potential data sparsity, and structural boundaries within the input feature space are rigorously identified prior to the initiation of advanced algorithmic training.

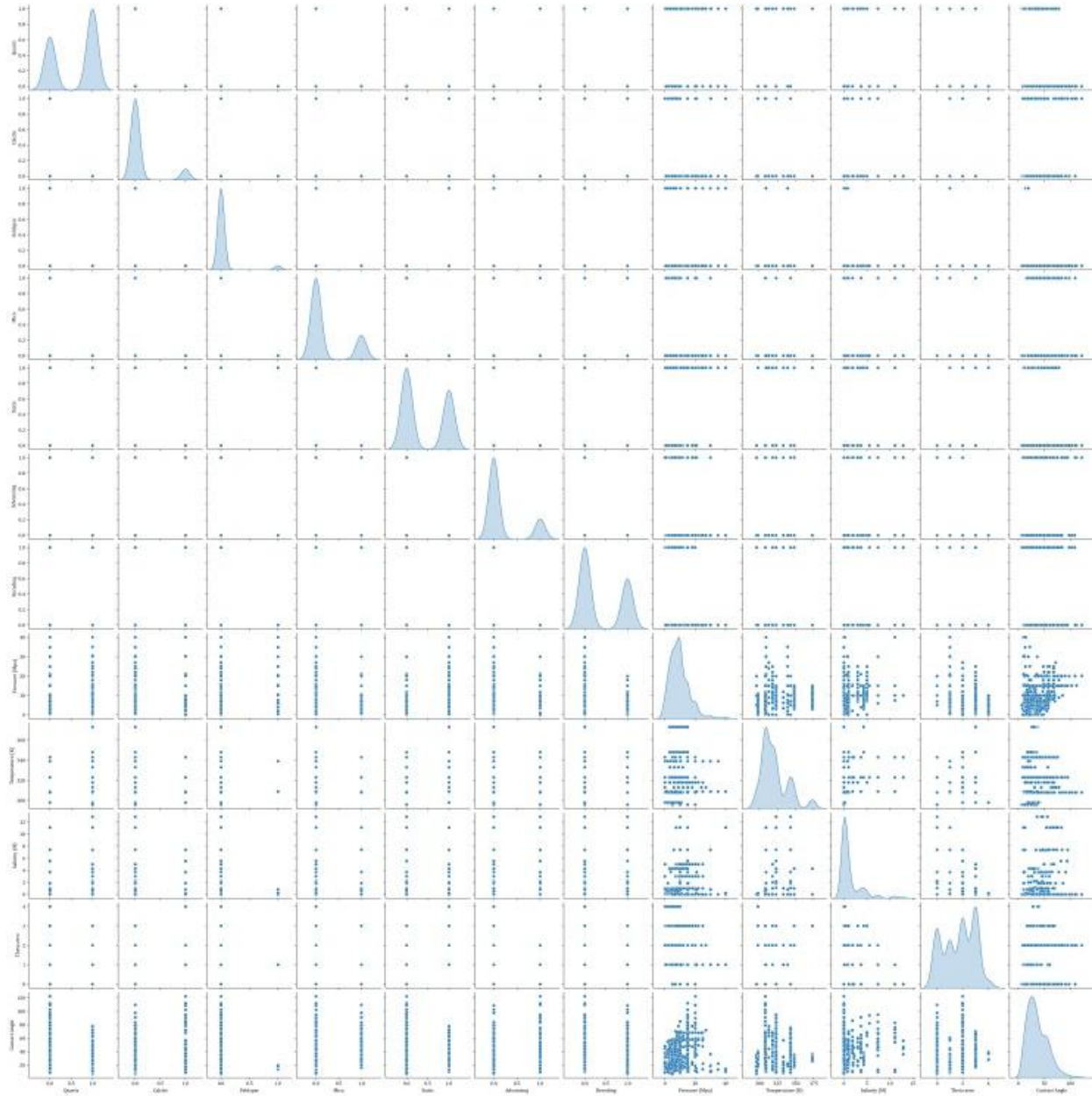


*Figure 2. Comprehensive scatter matrix plot illustrating the univariate kernel density distributions and bivariate correlations of the mineralogical, thermodynamic, and interfacial variables within the compiled dataset.*

In culmination, a highly robust and scientifically comprehensive foundation required for the subsequent execution of the advanced computational modeling framework is successfully established by the meticulously compiled and rigorously visualized dataset. By integrating a broad spectrum of critical thermodynamic, chemical, and mineralogical variables alongside carefully quantified baseline interfacial metrics, the complex multidimensional nature of CO<sub>2</sub> wettability is accurately captured for analytical processing. This rigorously structured empirical matrix is subsequently utilized so that the predictive efficacy of the designated machine learning algorithms can be systematically trained, validated, and comparatively evaluated, thereby ensuring that multiphase interfacial behaviors crucial for optimizing sustainable geological carbon sequestration operations can be reliably forecasted by the final mathematical models.

### **2.3 Cross-Validation Framework**

The compiled dataset consisting of the comprehensively gathered empirical observations was systematically partitioned into distinct training and testing subsets to ensure rigorous evaluation of the predictive models. Specifically, a standard allocation ratio was employed, wherein 80% of the data points were randomly assigned to the training subset, while the remaining 20% were strictly reserved for the testing subset [48, 49]. This rigorous isolation of the testing data is scientifically mandated so that the final generalization capability of the optimized machine learning algorithms can be objectively assessed on entirely unseen interfacial measurements, thereby inherently preventing any potential data leakage during the subsequent algorithmic training phases.

To facilitate an optimal hyperparameter tuning process while simultaneously mitigating the risk of model overfitting, a robust k-fold cross-validation methodology was strictly applied exclusively

to the isolated training subset. In this methodological framework, the  $k$ parameter was mathematically defined as 5, meaning the training dataset was effectively subdivided into five mutually exclusive and approximately equal-sized partitions, commonly referred to as individual folds. During each independent iteration of the cross-validation procedure, four of these distinct folds were systematically aggregated to serve as the immediate training corpus, whereas the single remaining fold was temporarily held out to function as a localized validation set.

The aforementioned five-fold iteration process was systematically repeated exactly five times, ensuring that every single partition within the designated training subset was utilized precisely once as the localized validation fold. Within this iterative structural framework, an exhaustive Grid Search strategy was heavily implemented so that the multidimensional hyperparameter space of each respective machine learning algorithm could be meticulously explored and rigorously evaluated. By systematically traversing a predefined grid of algorithmic configurations, the specific combination of hyperparameters that yielded the highest predictive accuracy across all five localized validation folds was mathematically identified and officially selected for the final model construction.

Upon the successful completion of the grid search-driven cross-validation process, the mathematically optimal hyperparameter configurations were subsequently extracted and applied to retrain the respective machine learning models using the entirety of the 80% training subset. This comprehensive retraining phase is fundamentally essential, as the maximum available data volume within the training domain is leveraged to mathematically solidify the complex correlational mappings between the geochemical inputs and the contact angle output. Ultimately, the strictly isolated 20% testing subset, which remained entirely untouched throughout the extensive hyperparameter optimization iterations, was subsequently deployed so that the

definitive, unbiased predictive performance of the fully tuned computational models could be rigorously quantified.

The quantitative efficacy of the developed algorithms was strictly evaluated employing a comprehensive suite of statistical metrics, specifically including the Mean Squared Error (MSE), the Coefficient of Determination ( $R^2$ ), the MRD%, and the Relative Error (RE%). The MSE fundamentally measures the average squared difference between the estimated predictions and actual experimental values, while the  $R^2$  metric statistically quantifies the proportion of variance in the dependent variable that is mathematically predictable from the independent variables. Furthermore, AARE% and RE% are mathematically utilized to assess the proportional deviation of the predicted outputs relative to the observed experimental data across the respective scales. To ensure absolute scientific transparency regarding algorithmic generalization capabilities, these statistical metrics are reported entirely separately for the training and testing subsets, whereas the overarching total performance metric for each distinct model is systematically calculated and documented as a mathematically weighted average of the respective subset evaluations.

## **2.4 Statistical Outlier Detection Framework**

To rigorously evaluate the statistical distribution and underlying variability of the compiled dataset, a comprehensive statistical analysis was conducted utilizing Monte Carlo-based outlier detection alongside standard boxplot visualizations, as presented in Figure 3. The implemented Monte Carlo simulation established defining confidence intervals bounded by  $\pm 2$  standard deviations from the dataset mean, visually represented by the distinct dashed lines within the generated frequency histogram [50, 51]. The histogram explicitly reveals a predominantly right-skewed distribution of the measured values, while the complementary boxplot systematically identifies the presence of several discrete data points residing significantly beyond the uppermost

interquartile range whisker. These statistical techniques collectively serve as a robust mechanism for mapping the structural spread of the experimental data and pinpointing numerical anomalies within the high-value spectrum of the targeted interfacial parameters.

Despite the mathematical identification of statistical outliers residing above the established upper threshold, a deliberate and scientifically motivated decision was strictly made to retain all initial data points within the final computational workflow. The identified deviations are not interpreted as erroneous experimental measurements; rather, they legitimately represent extreme, yet physically realistic, thermodynamic and geochemical conditions inherently present within heterogeneous geological formations. Consequently, the isolated outlier detection procedure was utilized exclusively for analytical cross-checking and overarching data integrity verification purposes. By entirely avoiding the arbitrary elimination of extreme empirical values, the fundamental generalizability of the developed machine learning algorithms is thoroughly preserved, thereby ensuring that the ultimate predictive framework remains highly robust and directly applicable across the complete spectrum of conceivable real-world carbon capture and storage environments.

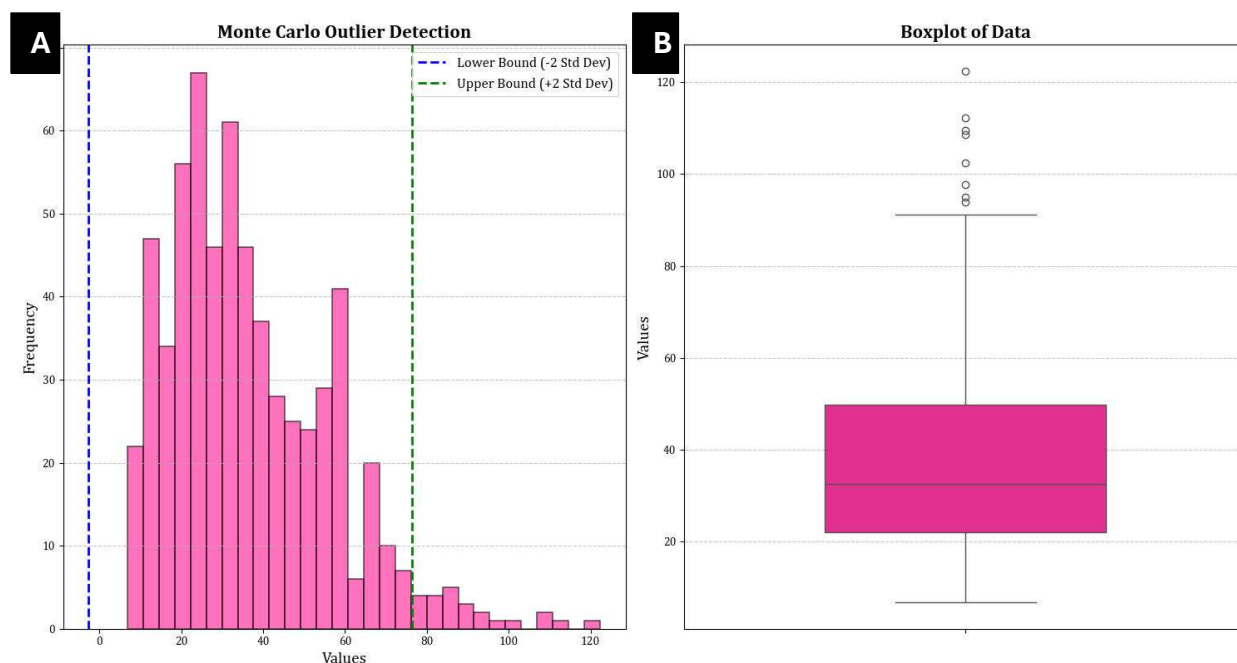


Figure 3. Monte Carlo outlier detection histogram and standard boxplot visualization applied to the experimental dataset

## 2.5 Feature Pearson Correlation Analysis

To systematically quantify the linear interdependencies among the diverse input features and the target contact angle, a comprehensive Pearson correlation matrix was computationally generated and visualized as a color-coded heatmap, detailed in Figure 4. An in-depth examination of the respective mineralogical descriptors reveals that the weight fraction of Mica exhibits the strongest positive correlation ( $r=0.39$ ) with the measured contact angle, indicating a distinct mineralogical transition towards CO<sub>2</sub>-wetness. This specific physical phenomenon is scientifically attributed to the highly layered silicate structure and the specific surface charge distribution inherent to mica, which physically facilitates enhanced CO<sub>2</sub> adsorption in deep saline environments. Conversely, Quartz demonstrates a notable negative correlation ( $r=-0.33$ ), inherently signifying its strong affinity for maintaining strict water-wetness (hydrophilicity) under analogous reservoir conditions due to its highly stable crystalline silica network and robust hydrogen-bonding capabilities with the surrounding aqueous phase.

Furthermore, the comprehensive correlation analysis mathematically elucidates the significant thermodynamic influence of the varied operational parameters, specifically highlighting definitive positive correlations for both the system pressure ( $r=0.25$ ) and the brine salinity ( $r=0.25$ ) with the resulting macroscopic contact angle. From a fundamental physicochemical perspective, an isothermal escalation in ambient pressure directly increases the mass density of the injected CO<sub>2</sub> phase, thereby intensifying the intermolecular van der Waals interactions at the fluid-mineral interface and consequently promoting a progressively more CO<sub>2</sub>-wet state. Simultaneously, elevated ionic salinity levels thermodynamically induce a critical compression of the electrical double layer surrounding the electrically charged mineral surfaces; this crucial electrostatic suppression critically diminishes the inherent repulsive hydration forces between the brine and the rock matrix, thereby allowing the non-polar CO<sub>2</sub> molecules to more effectively displace the resident intergranular water film.

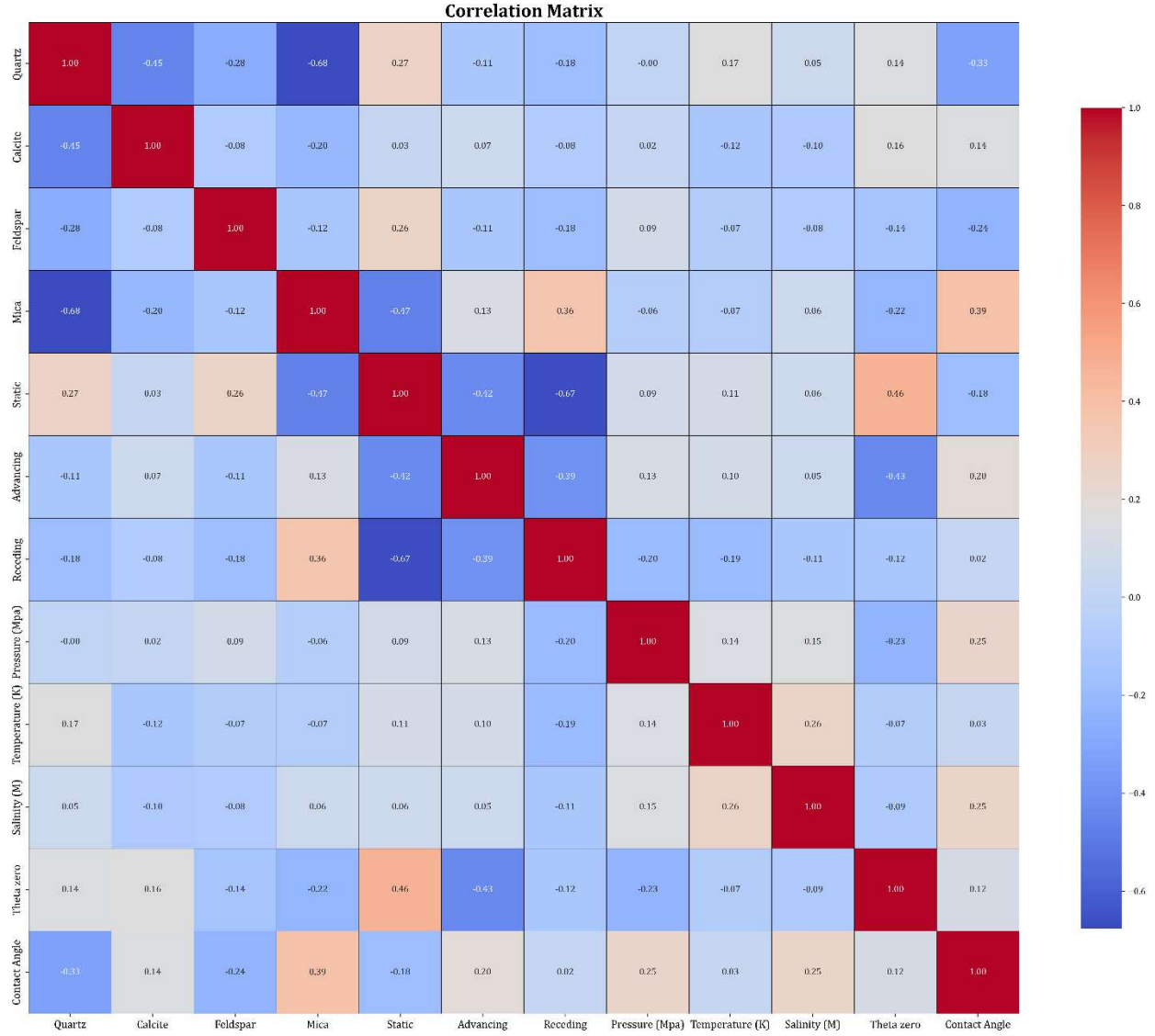


Figure 4. Pearson correlation matrix heatmap illustrating the pairwise linear relationships between geochemical descriptors, thermodynamic conditions, and interfacial contact angles

### 3. Results and discussions

#### 3.1 Predictive Algorithm Evaluation

The quantitative assessment of the fifteen deployed machine learning algorithms is systematically documented in Table 1, focusing specifically on the coefficient of  $R^2$ , MSE, and MRD% across the training, validation, and testing phases. A detailed comparative analysis clearly demonstrates that fundamental linear regression methodologies experience substantial algorithmic underfitting,

fundamentally failing to capture the complex, non-linear thermodynamic interactions inherently present within the multiphase system. Conversely, highly complex architectures such as Decision Trees exhibit severe computational overfitting; this instability is mathematically evidenced by a near-perfect training  $R^2$  of 0.9983 that precipitously collapses to a significantly degraded testing  $R^2$  of 0.7446. Consequently, advanced machine learning techniques must be carefully calibrated to balance these underlying mathematical extremes.

Based upon the rigorous statistical evaluation detailed in the compiled performance matrix, the CatBoost architecture is definitively established as the optimum predictive algorithm for mathematically modeling  $\text{CO}_2$  wettability behavior. This specific ensemble model achieves the absolute highest overarching performance during the critical testing phase, characterized by an exceptional testing  $R^2$  of 0.9357 alongside a significantly minimized testing MSE of 20.1358. Furthermore, the remarkably narrow performance differential observed between the initial training phase ( $R^2$  of 0.9877) and the final testing phase mathematically confirms that the CatBoost configuration successfully maintains the least computational overfitting among all evaluated high-dimensional algorithms. This superior structural generalization inherently ensures robust and highly reliable contact angle predictions across previously unseen geochemical reservoir conditions.

*Table 1. Comprehensive statistical performance matrix of the evaluated machine learning algorithms across the partitioned training, validation, and testing datasets*

| Model             | Train $R^2$ | Validation $R^2$ | Test $R^2$ | Train MSE | Validation MSE | Test MSE | Train MRD% | Validation MRD% | Test MRD% |
|-------------------|-------------|------------------|------------|-----------|----------------|----------|------------|-----------------|-----------|
| Linear Regression | 0.4414      | 0.3719           | 0.4765     | 233.7269  | 202.6346       | 163.9380 | 31.54      | 31.49           | 24.74     |
| Ridge Regression  | 0.4452      | 0.3922           | 0.4802     | 232.1249  | 196.0656       | 162.7849 | 30.84      | 30.40           | 24.82     |
| Lasso Regression  | 0.4473      | 0.3974           | 0.4846     | 231.2539  | 194.3904       | 161.4200 | 31.57      | 31.59           | 24.98     |
| Elastic Net       | 0.4443      | 0.3912           | 0.4792     | 232.5032  | 196.3903       | 163.1066 | 30.69      | 30.12           | 24.81     |

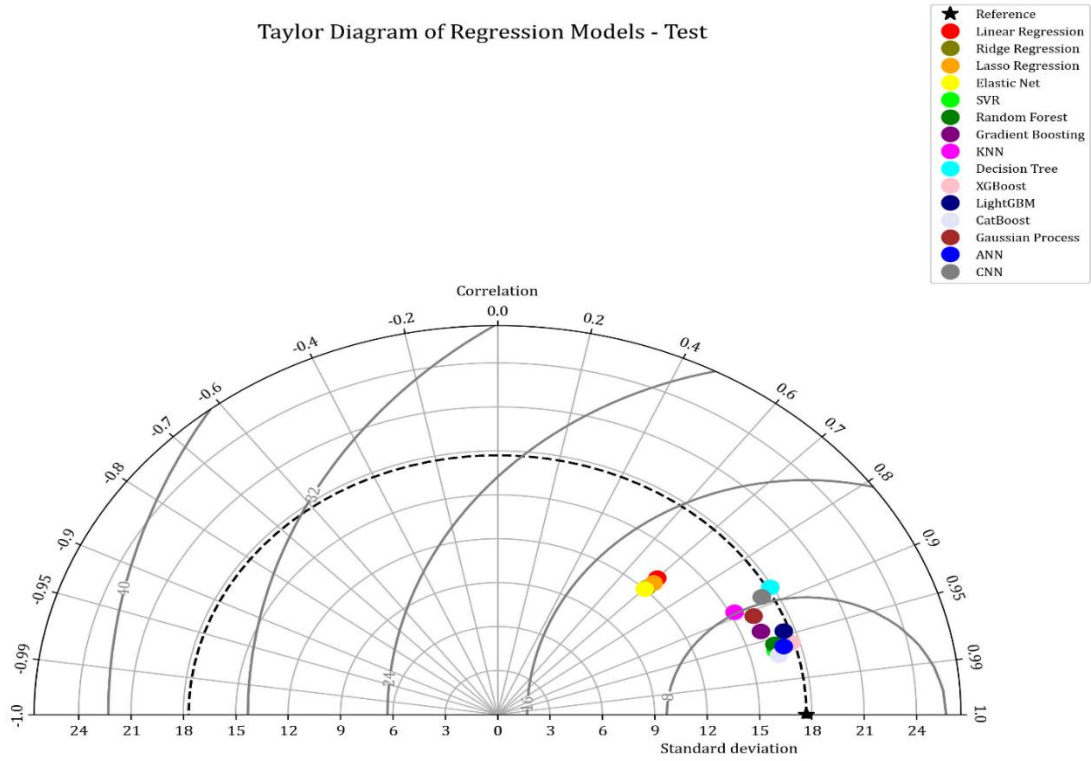
|                   |        |        |        |         |         |         |       |        |       |
|-------------------|--------|--------|--------|---------|---------|---------|-------|--------|-------|
| SVR               | 0.8809 | 0.8298 | 0.9286 | 49.8415 | 54.9148 | 22.3642 | 11.24 | 14.32  | 10.19 |
| Random Forest     | 0.9803 | 0.8940 | 0.9069 | 8.2391  | 34.1859 | 29.1457 | 5.61  | 12.30  | 10.19 |
| Gradient Boosting | 0.9241 | 0.8561 | 0.8658 | 31.7712 | 46.4255 | 42.0236 | 12.50 | 15.65  | 14.71 |
| KNN               | 0.8700 | 0.6938 | 0.7850 | 54.3811 | 98.7918 | 67.3267 | 12.55 | 15.82  | 13.71 |
| Decision Tree     | 0.9983 | 0.8516 | 0.7446 | 0.7124  | 47.8871 | 79.9734 | 0.22  | 15.78  | 15.34 |
| XGBoost           | 0.9980 | 0.9199 | 0.9156 | 0.8477  | 25.8373 | 26.4368 | 1.08  | 11.00  | 10.28 |
| LightGBM          | 0.9517 | 0.8399 | 0.8840 | 20.2083 | 51.6582 | 36.3169 | 11.59 | 22.52  | 14.68 |
| CatBoost          | 0.9877 | 0.9418 | 0.9357 | 5.1460  | 18.7730 | 20.1358 | 5.66  | 11.79  | 10.20 |
| Gaussian Process  | 0.9778 | 0.7756 | 0.8218 | 9.3047  | 72.3773 | 55.8089 | 6.01  | 19.29  | 13.55 |
| ANN               | 0.9884 | 0.8948 | 0.9231 | 4.8427  | 33.9515 | 24.0689 | 3.81  | 15.52  | 10.89 |
| CNN               | 0.9246 | 0.7380 | 0.7590 | 31.5275 | 84.5137 | 75.4710 | 19.49 | 118.06 | 48.05 |

### 3.2 Taylor Diagram Performance Evaluation

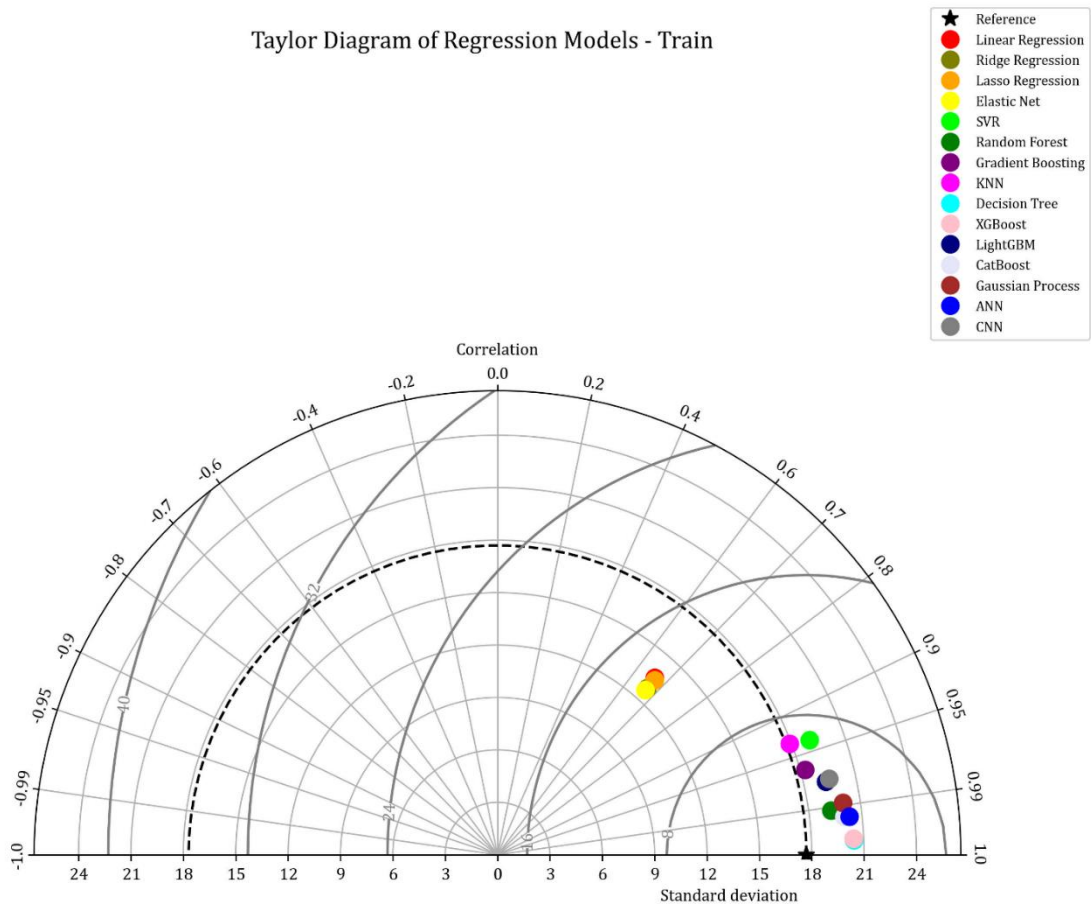
The comprehensive predictive accuracy and spatial concordance of the implemented machine learning algorithms are graphically evaluated utilizing Taylor diagrams, as presented in Figure 5 for the training, validation, and testing phases. This advanced multidimensional visualization framework mathematically integrates the Pearson correlation coefficient, the standard deviation, and the centered root mean square difference into a singular two-dimensional polar coordinate space. An analytical assessment of the training and validation distributions reveals that linear regression formulations, including LR, Ridge, Lasso, and EN, consistently exhibit pronounced spatial deviation from the reference observation point, fundamentally indicating restricted capability in capturing the inherent thermodynamic variance. Conversely, tree-based ensemble architectures and complex neural network structures demonstrate exceptional geometric proximity to the empirical reference node, signifying high correlational accuracy and optimal standard deviation alignment during the initial computational fitting phases.

A rigorous examination of the testing phase Taylor diagram definitively substantiates the superior generalization capacity of the advanced gradient boosting frameworks when applied to unseen thermodynamic and geochemical conditions. The CatBoost and XGBoost algorithms, alongside the ANN architecture, are geometrically positioned at the minimal radial distance from the experimental reference marker, mathematically denoting the highest positive correlation and the lowest centered root mean square error among all evaluated methodologies. Furthermore, while the DT algorithm demonstrates near-perfect structural alignment during the training phase, its noticeable displacement within the testing visualization graphically corroborates the severe computational overfitting previously identified during the numerical metric analysis. Ultimately, the CatBoost algorithm emerges as the optimal predictive framework, characterized by a spatial standard deviation virtually identical to the empirical dataset and a maximized correlation coefficient, thereby confirming its unparalleled robustness in modeling complex CO<sub>2</sub> wettability behavior.

Taylor Diagram of Regression Models - Test



Taylor Diagram of Regression Models - Train



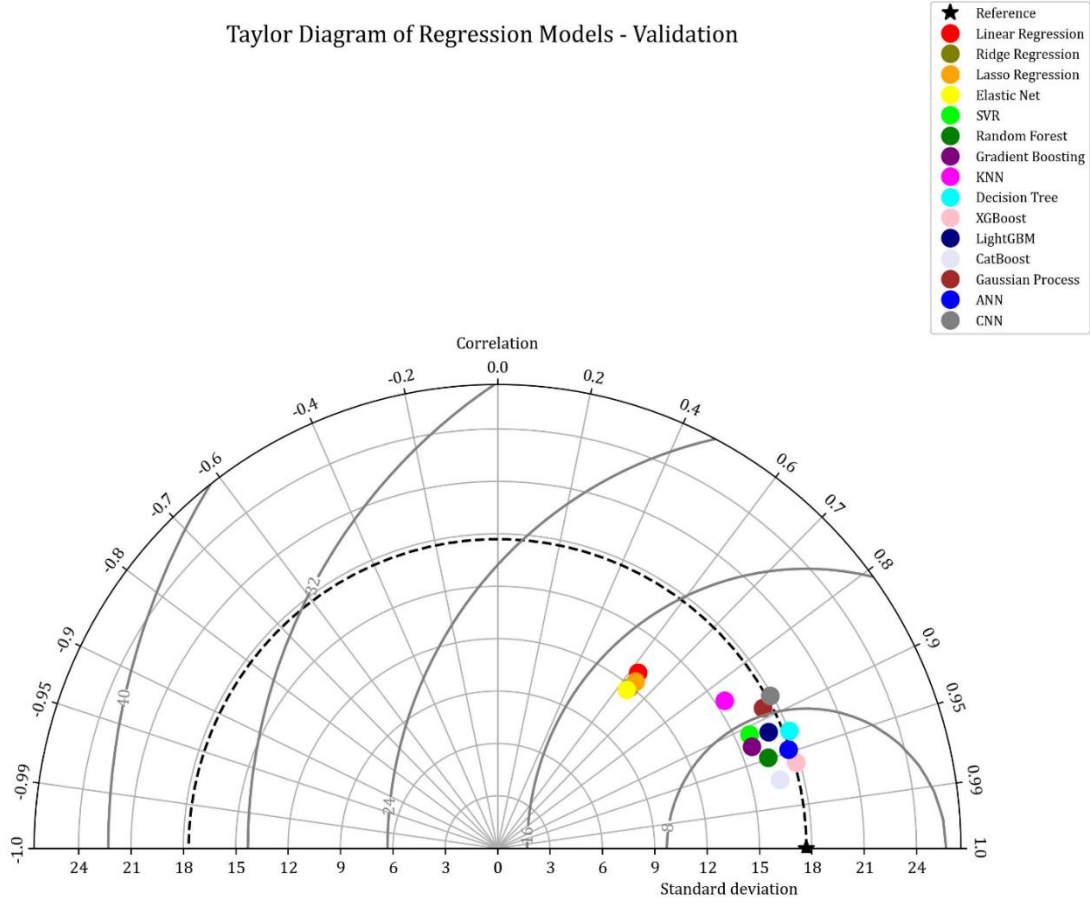


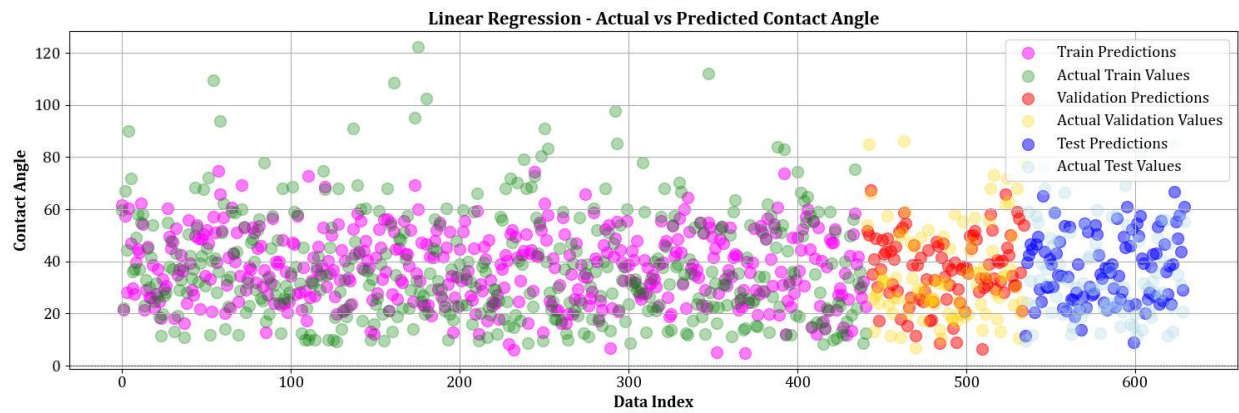
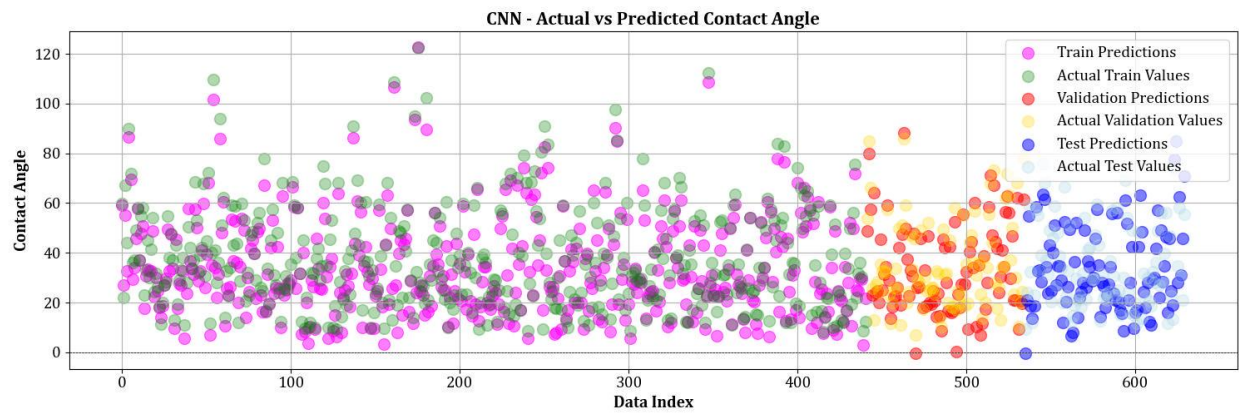
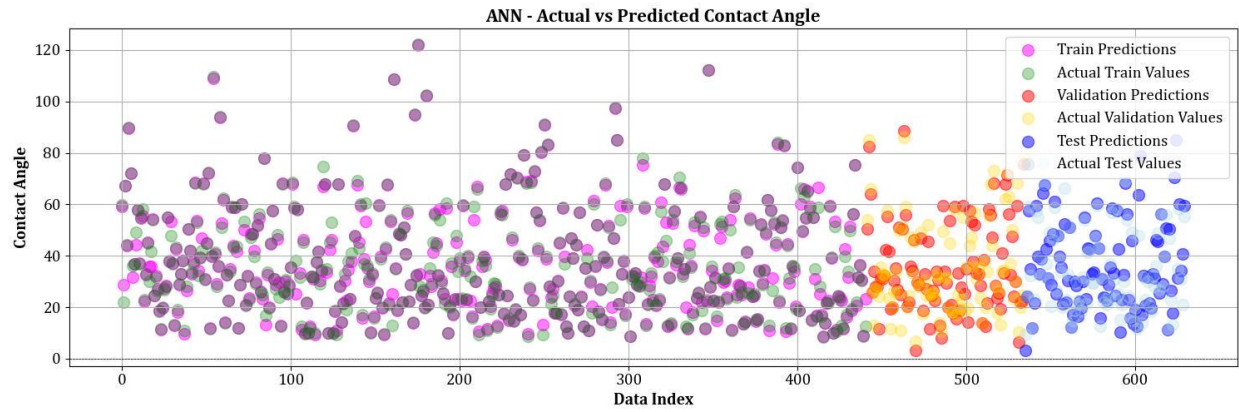
Figure 5. Taylor diagrams illustrating the multidimensional predictive performance of the evaluated regression models across the training, validation, and testing subsets, quantified through Pearson correlation, standard deviation, and centered root mean square difference

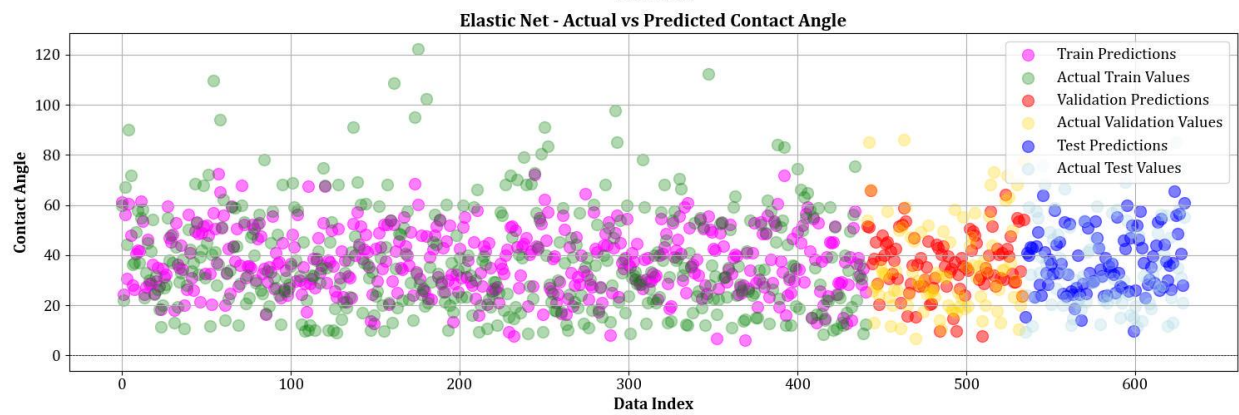
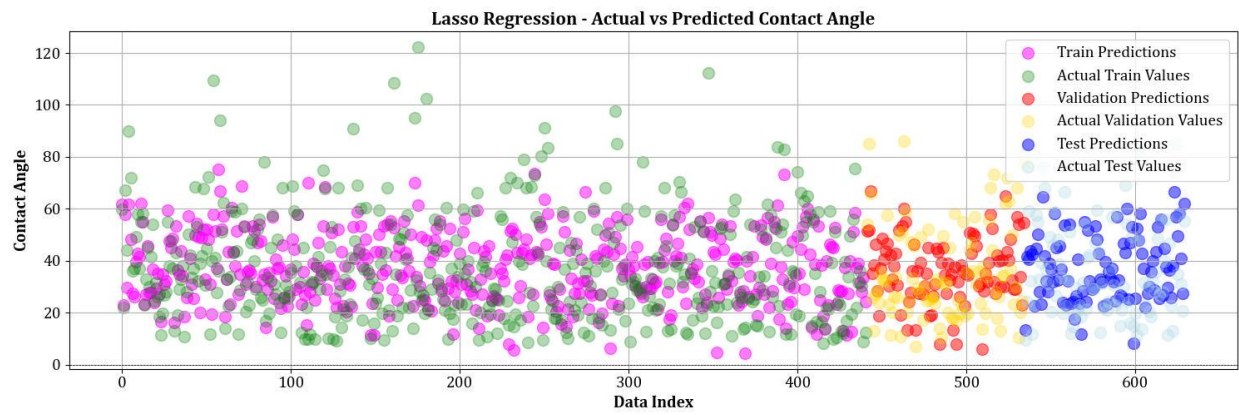
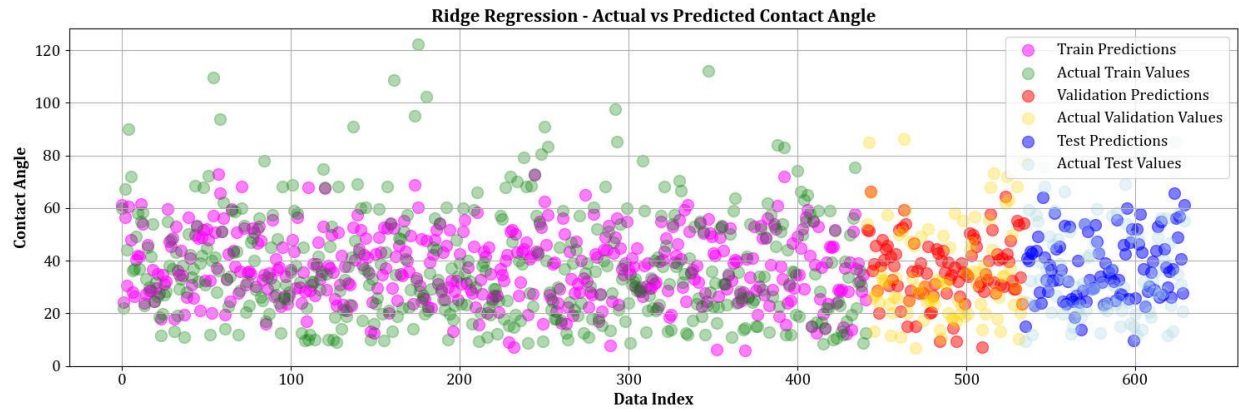
### 3.3 Visual Diagnostics and Error Distribution Analysis

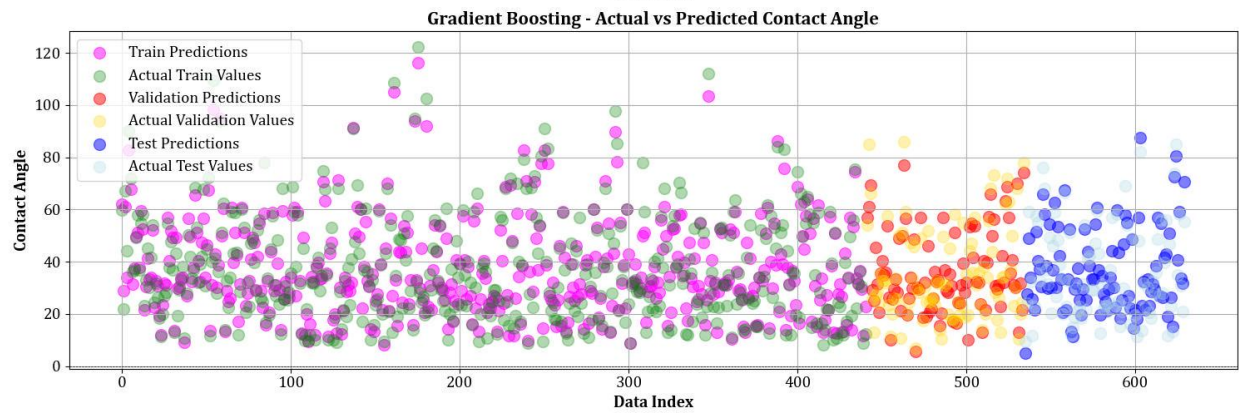
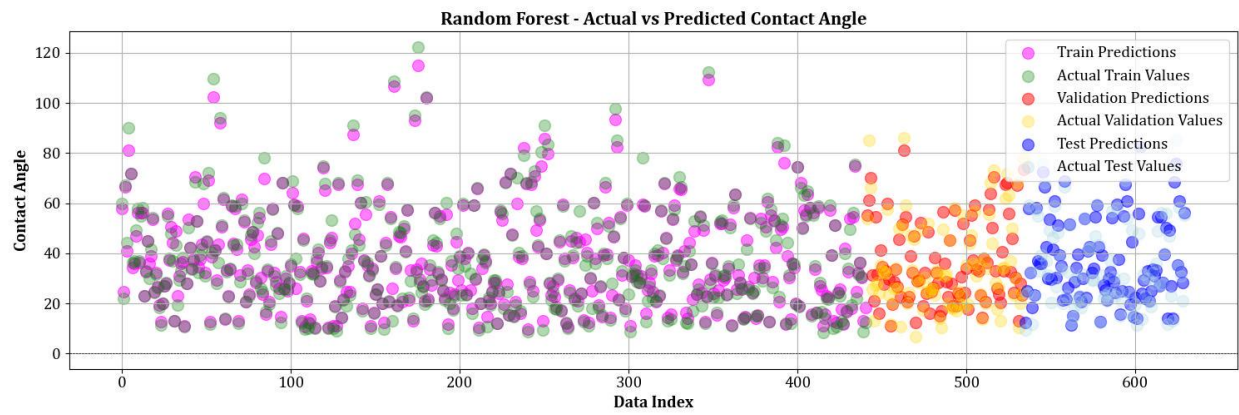
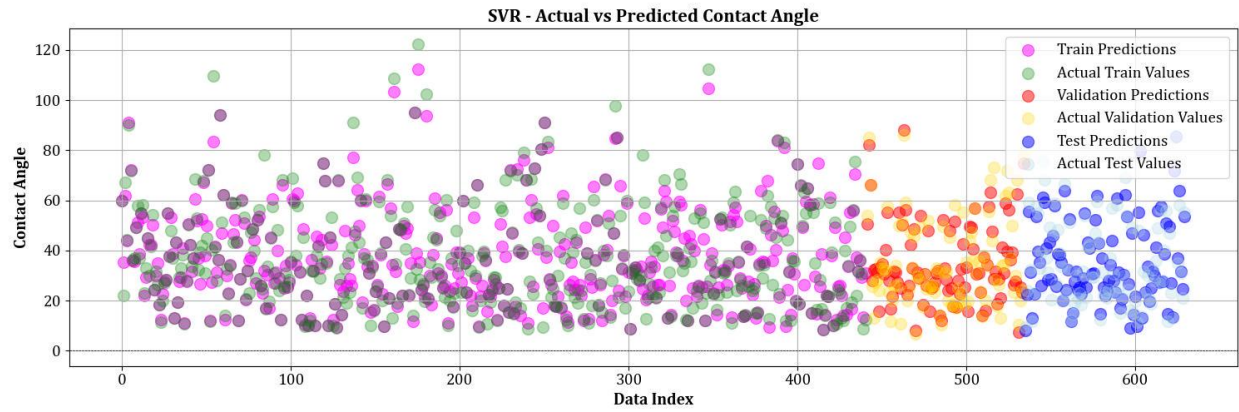
The index-based actual versus predicted contact angle visualization, presented in Figure 6, provides a rigorous, granular assessment of the algorithmic tracking capability across the sequential dataset indices for the training, validation, and testing subsets. An analytical review of these phase-specific alignments reveals that the foundational linear mathematical models, encompassing LR, Ridge, Lasso, and EN, exhibit significant structural deviations from the empirical targets. These formulations consistently fail to capture extreme contact angle fluctuations and localized non-linear transitions, generating smoothed predictive trajectories that

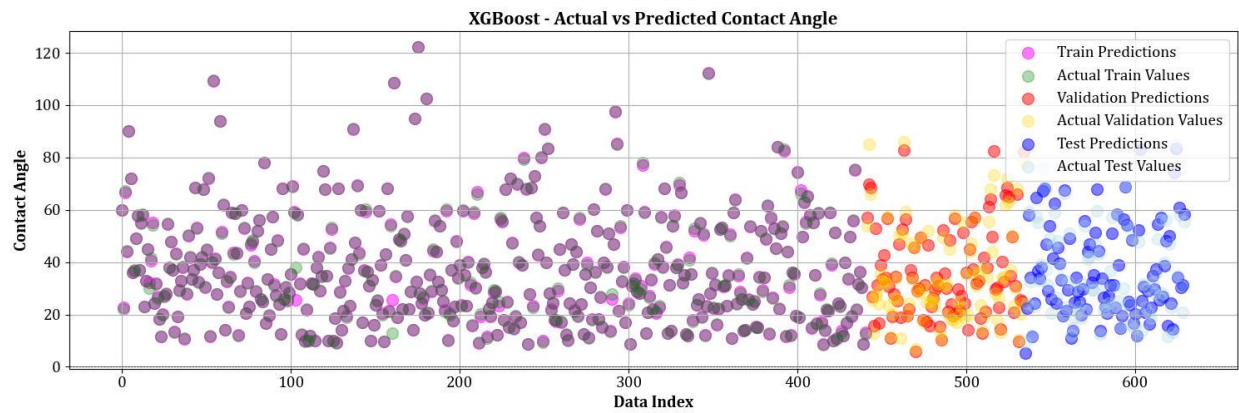
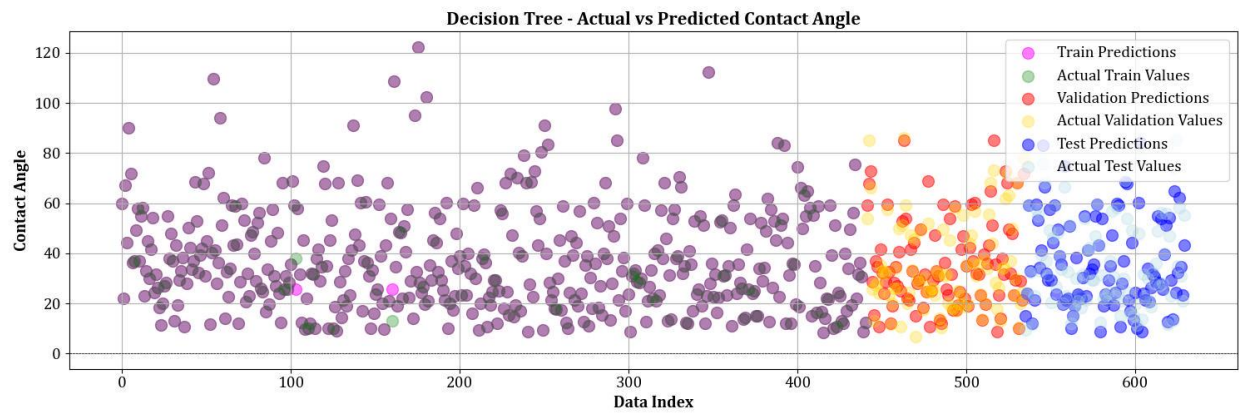
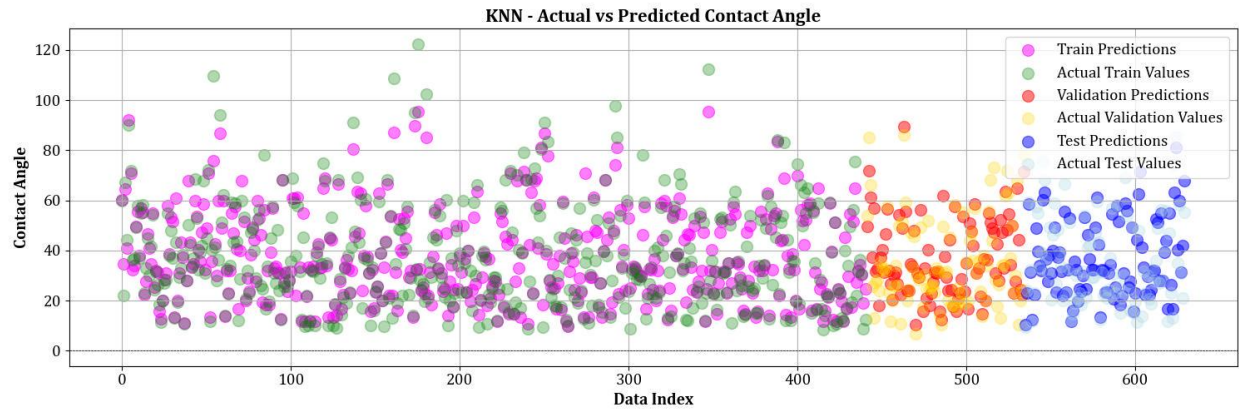
definitively confirm widespread algorithmic underfitting across the multidimensional thermodynamic space.

In stark contrast to the linear methodologies, the advanced tree-based ensemble frameworks, particularly XGBoost, LightGBM, and CatBoost, alongside the ANN architecture, are observed to tightly trace the actual empirical values across the entire index spectrum. While the DT algorithm demonstrates perfect predictive tracking during the training phase, its highly erratic deviations within the testing subset visually corroborate the severe structural overfitting diagnosed in earlier metric evaluations. Ultimately, the CatBoost algorithm exhibits the most consistent sequential alignment, successfully capturing both baseline trends and isolated wettability extremes without succumbing to the high-variance instability characterizing the weaker predictive models.









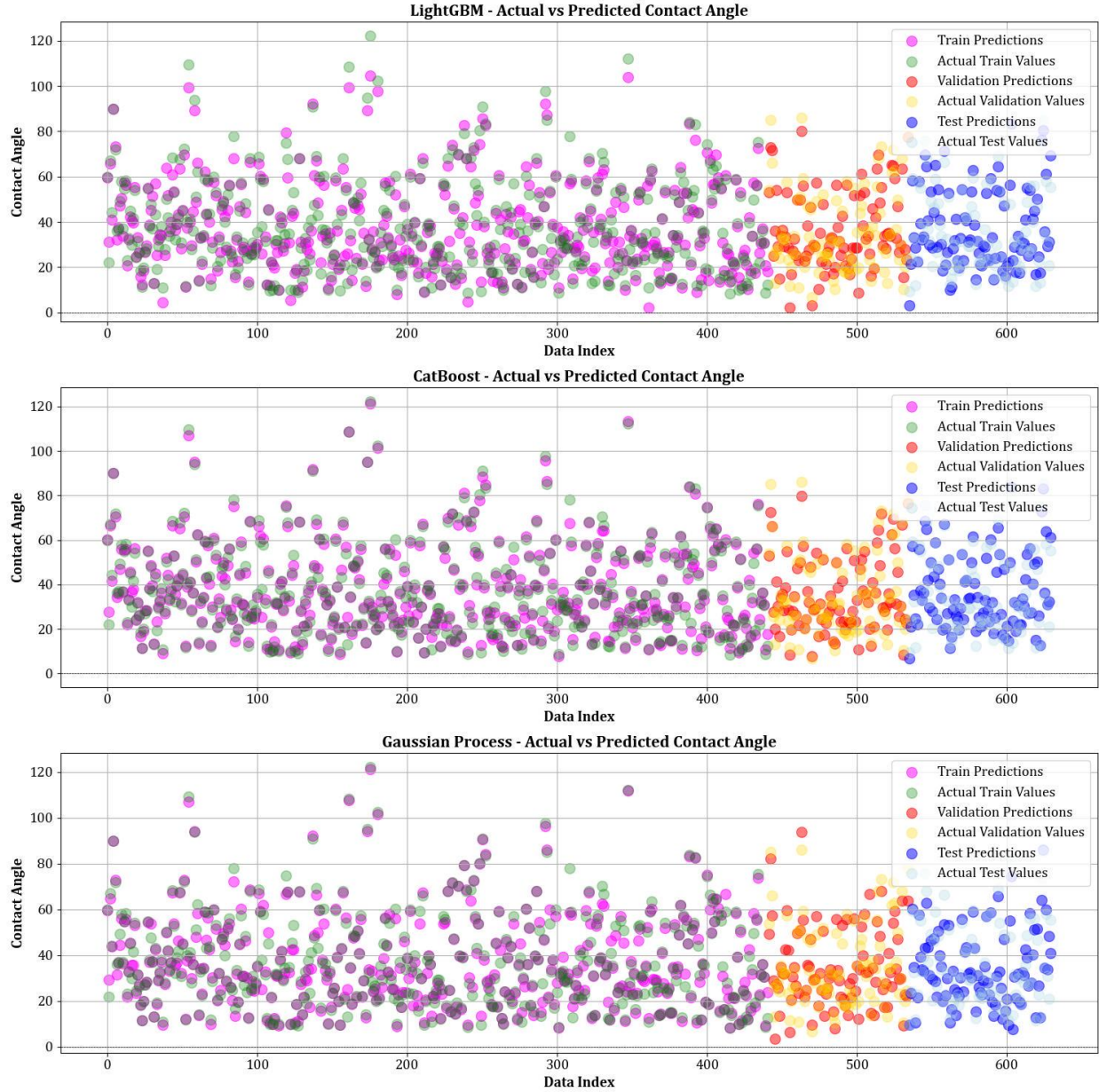


Figure 6. Sequential data index visualization of the actual versus algorithmically predicted CO<sub>2</sub> contact angles across the training, validation, and testing subsets for the evaluated machine learning models

The predictive goodness-of-fit and proportional algorithmic bias are further evaluated utilizing the parity scatter plots detailed in Figure 7, which map the predicted variables against the actual experimental values relative to the ideal  $y=x$  bisector line. The visual dispersion topologies associated with the linear models display a highly diffuse, cloud-like distribution around the ideal parity line, an observation numerically corroborated by their structurally suppressed testing phase

$R^2 \approx 0.48$ . This extensive spatial variance demonstrates a fundamental computational inability to accurately model the complex, non-linear fluid-rock interactions governing the CO<sub>2</sub> contact angle behavior.

Conversely, the parity distributions generated by the superior machine learning architectures emphasize exceptional mathematical convergence and predictive reliability. The spatial data points for the RF, XGBoost, CatBoost, and ANN models are observed to tightly cluster along the parity bisector across all data splits. While the DT model exhibits a severe structural collapse, transitioning from an artificial perfect alignment characterized by a training  $R^2$  of 1.00 to significant scatter with a testing  $R^2$  of 0.74, the CatBoost algorithm achieves an optimal geometric equilibrium. CatBoost maintains a tightly bounded dispersion profile, signifying minimal systematic error and high predictive fidelity across the full spectrum of contact angles.

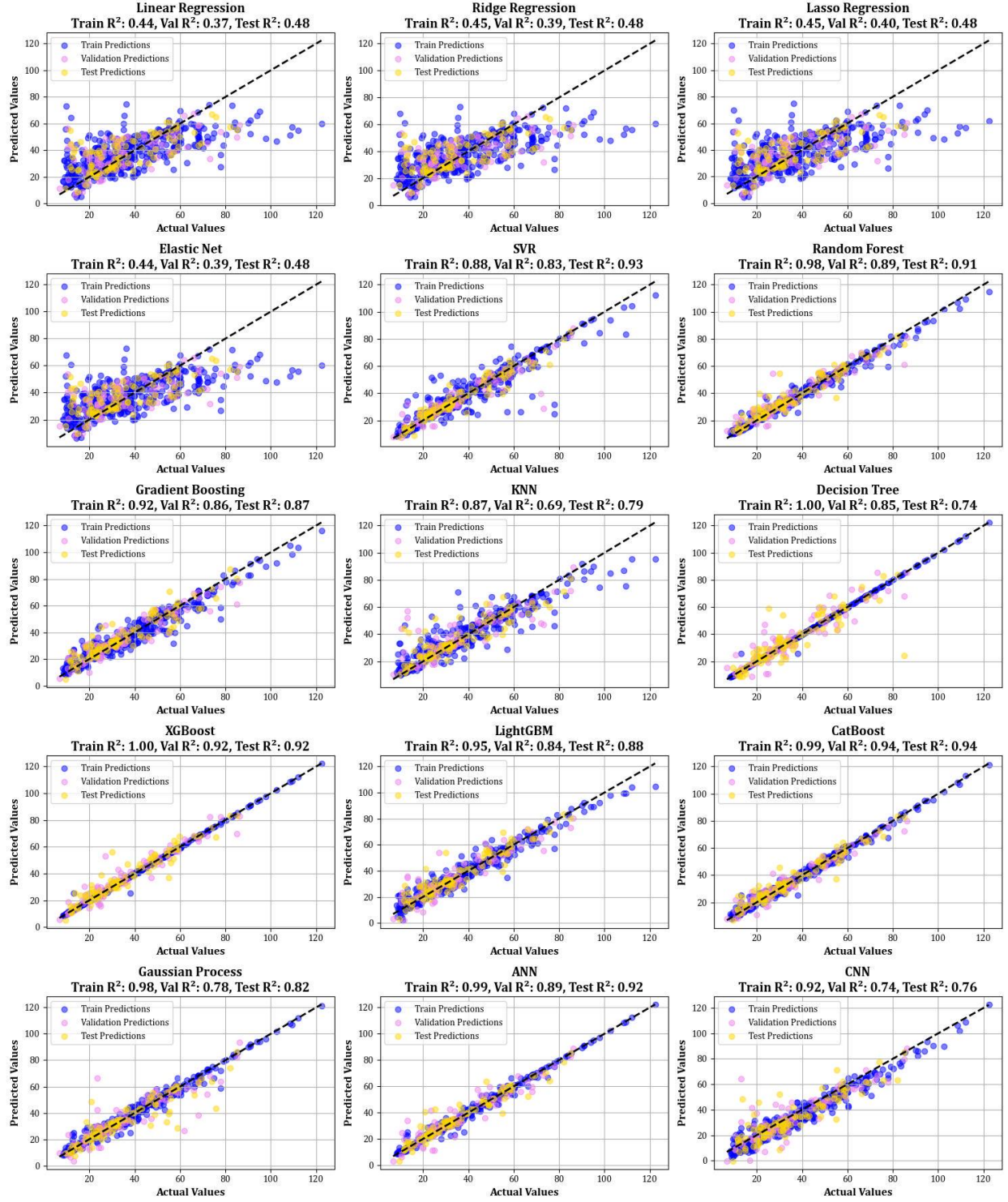


Figure 7. Parity scatter plots comparing the actual experimental contact angles against the algorithmically predicted values, featuring the ideal bisector and phase-specific  $R^2$  metrics

The magnitude of localized computational error is systematically interrogated through the relative deviation percentage profiles illustrated in Figure 8, which sequentially evaluate the predictive discrepancies across the dataset indices. The linear regression methodologies inherently generate highly volatile and unstable error profiles, frequently exhibiting massive relative deviations that significantly exceed acceptable engineering thresholds, particularly when predicting extreme or highly transitional wettability states. This intense error volatility definitively confirms that simplistic linear mathematical assumptions yield unacceptably wide margin of error boundaries when subjected to the inherent structural heterogeneities of CO<sub>2</sub>-brine-mineral geological interfaces.

When analyzing the advanced non-linear frameworks, the optimized gradient boosting algorithms, most notably CatBoost and XGBoost, successfully compress the relative deviation profiles into exceptionally narrow, low-magnitude bands that are heavily concentrated near the zero-error baseline. It is highly pertinent to observe that the CNN architecture demonstrates massive, isolated relative deviation spikes, likely attributable to localized algorithmic instability and matrix sparsity when applied to tabular thermodynamic datasets. However, the CatBoost framework consistently suppresses error propagation across the unseen testing subset, minimizing the RE metric and validating its deployment for high-precision geochemical modeling where unconstrained predictive deviations could compromise subsurface storage integrity.

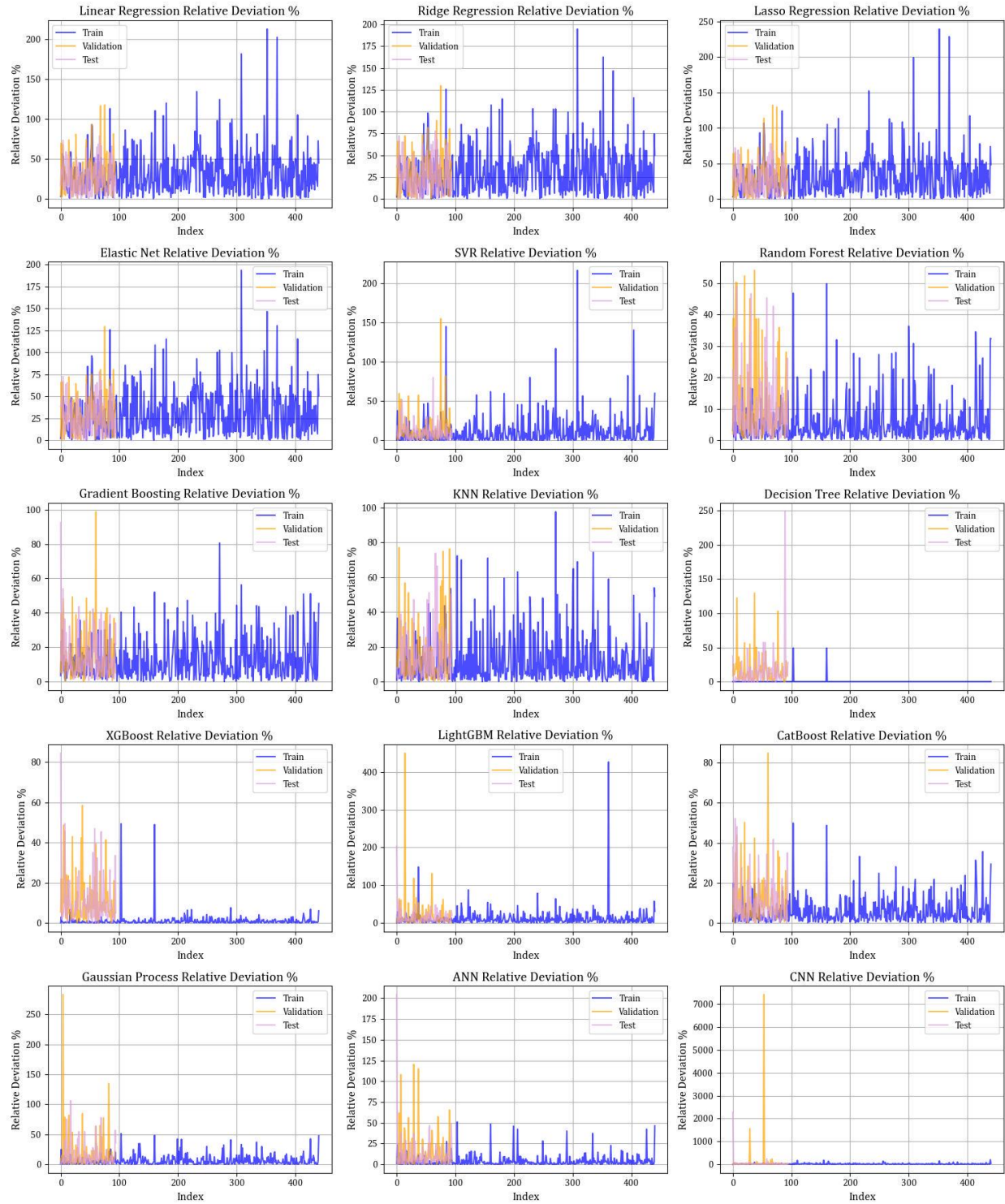


Figure 8. Sequential relative deviation percentage profiles evaluating the magnitude of predictive error across the dataset indices for the implemented machine learning regression models

The macroscopic statistical validity of the predictive methodologies is rigorously assessed through the frequency distribution histograms presented in Figure 9, which compare the algorithmically generated density topologies against the natural empirical distribution. A technically robust predictive framework must not only minimize discrete point-based metrics like MSE, but must also accurately replicate the macroscopic statistical behavior of the target variable. The linear algorithmic formulations are observed to generate structurally smoothed, highly centralized histograms that critically fail to capture the multi-modal and intrinsically skewed nature of the empirical wettability data, effectively homogenizing critical localized interfacial behaviors.

The advanced non-linear architectures mathematically reconstruct the complex frequency density topologies observed within the physical experimental domain with exceptional precision. The prediction histograms produced by the CatBoost, XGBoost, and ANN algorithms rigorously emulate the structural shape, central tendencies, and specific tail behaviors of the actual data across the training, validation, and testing subsets. The accurate preservation of these multi-modal distributional characteristics ensures that the probabilistic modeling of crucial intermediate-wet and strongly CO<sub>2</sub>-wet physical conditions remains statistically representative, further solidifying the CatBoost framework as a dominant computational tool for large-scale carbon capture and storage simulations.

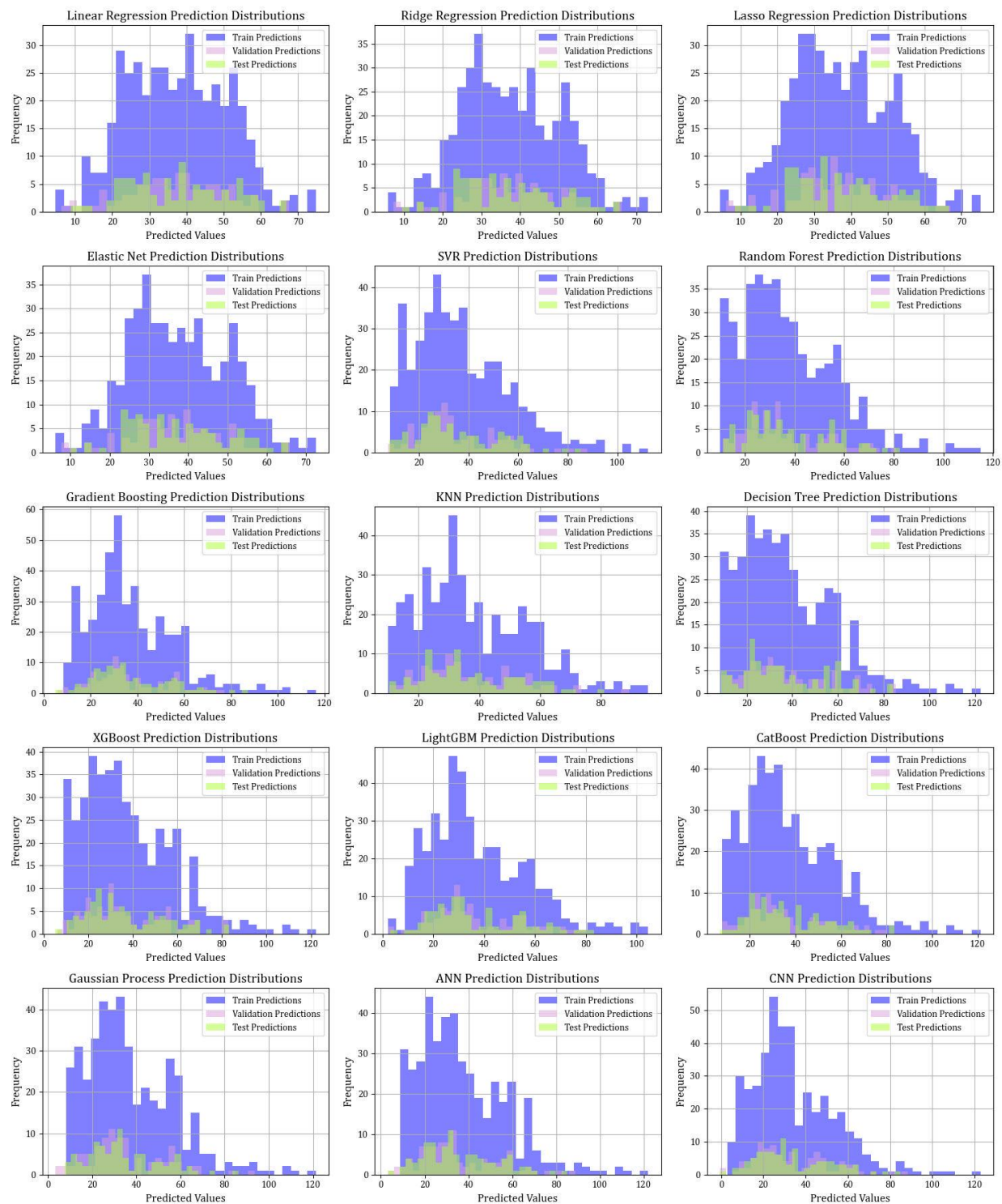


Figure 9. Frequency distribution histograms of the algorithmically predicted contact angles, illustrating the statistical replication of the physical variable density across the training, validation, and testing subsets

Integrating the comprehensive sequential tracking, parity dispersion, relative error magnitudes, and distributional replication analyses definitively solidifies the computational superiority of advanced gradient boosting frameworks over traditional linear and elementary distance-based models. This exhaustive visual and statistical diagnostic evaluation unequivocally designates the CatBoost algorithm as the optimal modeling paradigm for forecasting complex CO<sub>2</sub> wettability behaviors. By maximizing the R<sup>2</sup> coefficient, minimizing the MSE and RE metrics, and entirely bypassing the severe algorithmic overfitting trap mathematically demonstrated by the DT framework, CatBoost delivers a highly generalizable, mathematically stable architecture fundamentally essential for the sustainable engineering and quantitative risk assessment of deep geologic sequestration sites.

### **3.4 Residual Error Distribution and Density Analysis**

The statistical dispersion and localized variance of the predictive errors are rigorously quantified through the violin plot visualization of residuals, as illustrated in Figure 10. This advanced graphical technique mathematically integrates conventional boxplot metrics—specifically the median and interquartile ranges, with non-parametric kernel density estimation to provide a comprehensive topological representation of the error distribution for each evaluated algorithm. An analytical inspection of the residual distributions associated with the foundational linear models, namely LR, Ridge, Lasso, and EN, reveals significantly elongated and expanded geometries, indicating highly dispersed predictive errors. These distended density profiles conclusively demonstrate massive predictive uncertainty and fundamentally validate the structural inadequacy of purely linear mathematical functions in capturing the complex, non-linear thermodynamic fluctuations inherent to CO<sub>2</sub> wettability behavior.

Conversely, the residual distributions generated by the sophisticated tree-based ensemble architectures and deep learning frameworks exhibit highly optimized error concentration, predominantly constrained near the absolute zero-error baseline. The XGBoost and CatBoost algorithms, alongside the ANN formulation, display tightly compressed interquartile ranges and minimized density tails, mathematically denoting exceptionally low predictive variance and negligible systematic bias. It is highly pertinent to observe that while the DT algorithm demonstrates a centralized median, it simultaneously exhibits excessively prolonged outlier tails, a geometric anomaly that physically corroborates the severe algorithmic overfitting and testing instability previously identified. Ultimately, the CatBoost algorithm emerges with the most structurally compact residual distribution, ensuring maximum predictive reliability and minimizing the localized computational discrepancies that frequently compromise complex geochemical simulations.

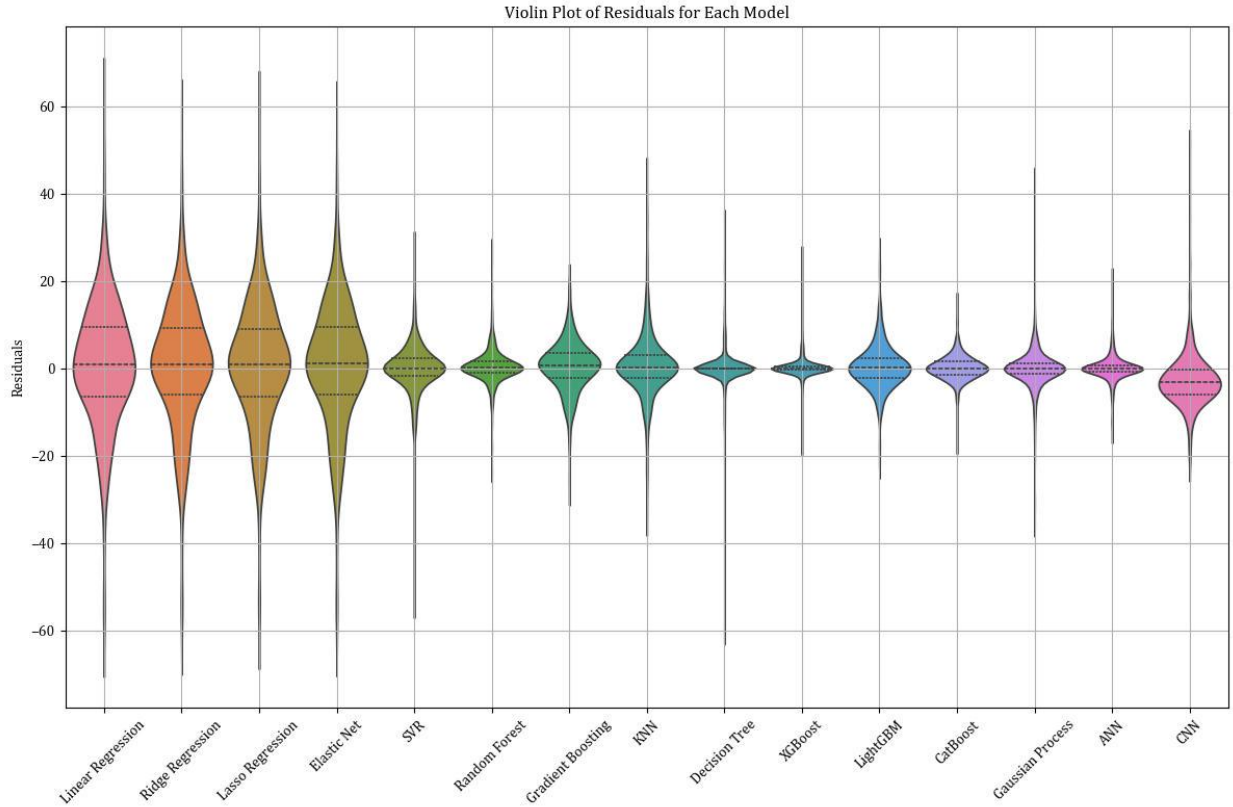


Figure 10. Violin plot visualization of the predictive residuals for the evaluated machine learning algorithms, integrating interquartile range statistics with non-parametric kernel density estimation to quantify error dispersion

The probabilistic frequency and centrality of the computational errors are further mathematically interrogated utilizing the residual density plot detailed in Figure 11, which maps the continuous probability density functions of the predictive deviations across all fifteen algorithms. Within this analytical framework, an optimal predictive model is theoretically characterized by a highly kurtotic, symmetrical peak centered perfectly upon the  $x=0$  residual axis, signifying a dominant probability of generating zero-error predictions. The macroscopic evaluation of the linear regression methodologies reveals characteristically broad, flattened, and highly diffuse density curves that span a substantial range of negative and positive error magnitudes. This significant spatial diffusion mathematically confirms that linear models systematically fail to minimize the residual variance, directly resulting in the elevated MSE and AARE metrics recorded during the numerical validation phase.

In direct contrast to the diffuse error probabilities of the linear formulations, the advanced gradient boosting algorithms generate sharply peaked, highly concentrated density topologies that closely approximate an ideal zero-error distribution. The XGBoost and CatBoost frameworks exhibit exceptionally prominent vertical density spikes precisely bisected by the zero-error reference line, indicating an overwhelming statistical probability of highly accurate contact angle predictions. While the CNN architecture displays a noticeably skewed and decentralized density profile, suggesting localized matrix instability or structural under-optimization, the CatBoost algorithm maintains a flawlessly symmetrical and maximized central peak. This mathematically superior residual density profile signifies optimal algorithmic calibration, ensuring that the selected predictive framework operates with minimal structural bias across the evaluated operational and mineralogical parameter space.

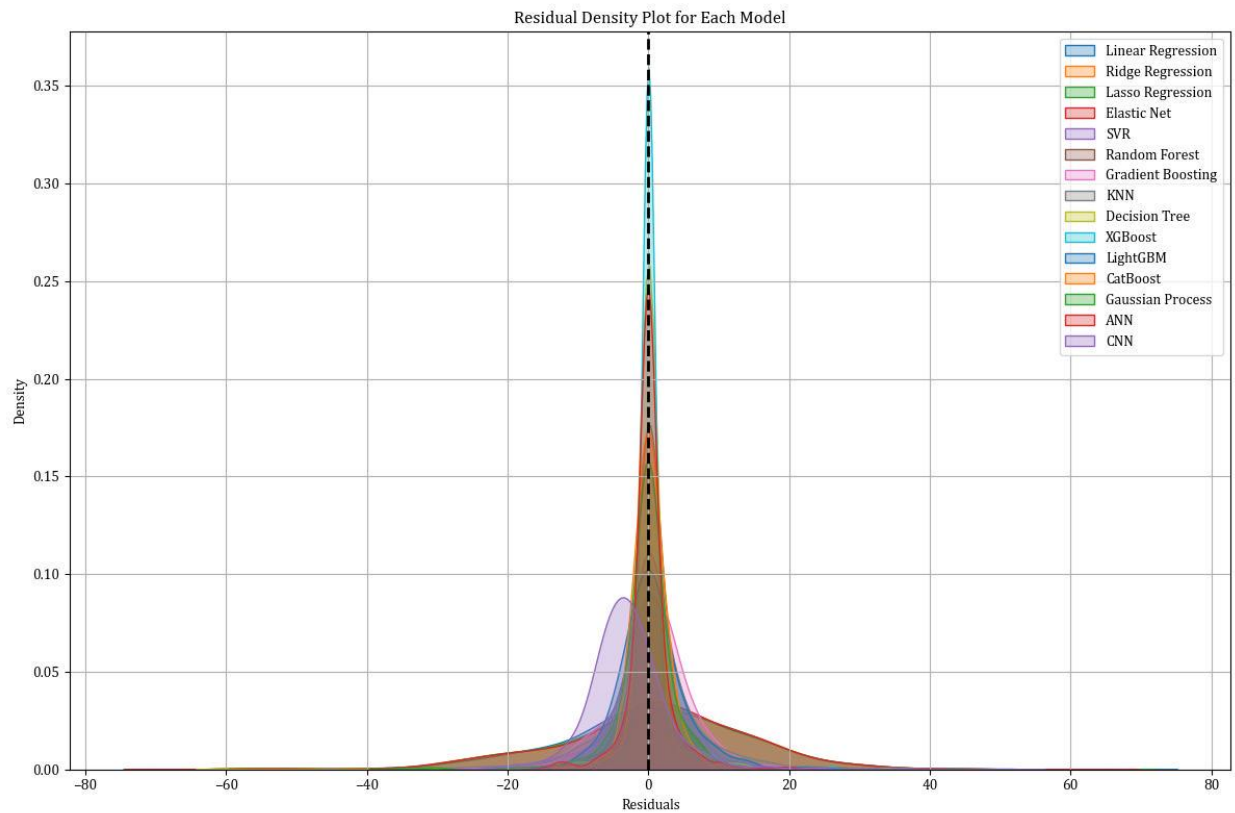


*Figure 11. Continuous residual probability density functions for the implemented machine learning regression models, illustrating the statistical frequency, centrality, and kurtosis of the predictive errors relative to the zero-error baseline*

Integrating the rigorous statistical insights derived from the violin plot dispersion metrics and the continuous probability density functions decisively concludes the error distribution analysis, reinforcing the unparalleled computational superiority of the CatBoost architecture. By mathematically demonstrating the tightest interquartile error constraint, the absolute minimization of extreme outlier deviations, and a highly maximized zero-error probability peak, the CatBoost algorithm proves completely immune to the extensive variance and systematic biases that critically degrade linear and unregularized tree-based models. This definitive stabilization of the predictive residuals verifies the robustness of the framework, confirming its suitability for deployment within high-stakes CO<sub>2</sub> sequestration engineering workflows, where minimizing quantitative uncertainty at the CO<sub>2</sub>/brine/mineral interface is strictly mandatory for the sustainable and secure design of deep geological carbon storage sites.

### **3.5 SHAP-Based Interpretability Analysis**

The determination of the most influential physical and mineralogical variables governing the contact angle at the CO<sub>2</sub>/brine/mineral interface is systematically conducted utilizing dual interpretability frameworks, as presented in Figure 12. Initially, the relative contribution of each input parameter toward reducing the computational variance is delineated by the intrinsic feature importance derived from RF algorithm. Based on the calculated importance scores, operational Pressure (MPa) is identified as the most dominant predictive parameter, exceeding an importance score of 0.30, which mathematically underscores the critical role of hydrostatic stress in dictating CO<sub>2</sub> density, interfacial tension, and dynamic wettability alterations. Furthermore, highly significant contributions are exhibited by the mineralogical presence of Mica and the intrinsic baseline angle,  $\theta_0$  (Theta zero), both registering importance scores approximating 0.16. It is

indicated by these metrics that the specific surface chemistries of phyllosilicates and initial wetting states fundamentally control the thermodynamic equilibrium of the multiphase system. Conversely, minimal importance scores approaching absolute zero are allocated to mineralogical parameters such as Feldspar and specific dynamic measurements like the Receding angle, suggesting a negligible independent influence on the final contact angle behavior within the modeled operational envelope.

To rigorously validate the algorithmic feature ranking and provide a model-agnostic quantification of variable impact, SHapley Additive exPlanations (SHAP) analysis is additionally implemented to evaluate the mean marginal contributions of each independent feature. The primary sensitivities previously identified by the tree-based methodology are robustly corroborated by the mean SHAP values, visualized in the adjacent subplot of Figure 12, while a refined hierarchical ordering based on absolute predictive magnitude is simultaneously provided. Within this cooperative game-theoretic framework, Theta zero is identified as the paramount driving factor, registering a peak mean SHAP value exceeding 8.0, thereby confirming that the predefined static boundary condition severely constraints the algorithmic output. Subsequently, a profound predictive influence is consistently demonstrated by Mica and Pressure (MPa), yielding mean SHAP values of approximately 7.5 and 7.2, respectively, which physically validates the intense dependence of CO<sub>2</sub> wettability on localized mineral-fluid interactions and injection-induced reservoir stresses. Furthermore, moderate SHAP sensitivities are maintained by Temperature (K) and Salinity (M), reflecting their essential but secondary roles in modifying brine ionic strength and aqueous fluid viscosity during structural carbon trapping processes.

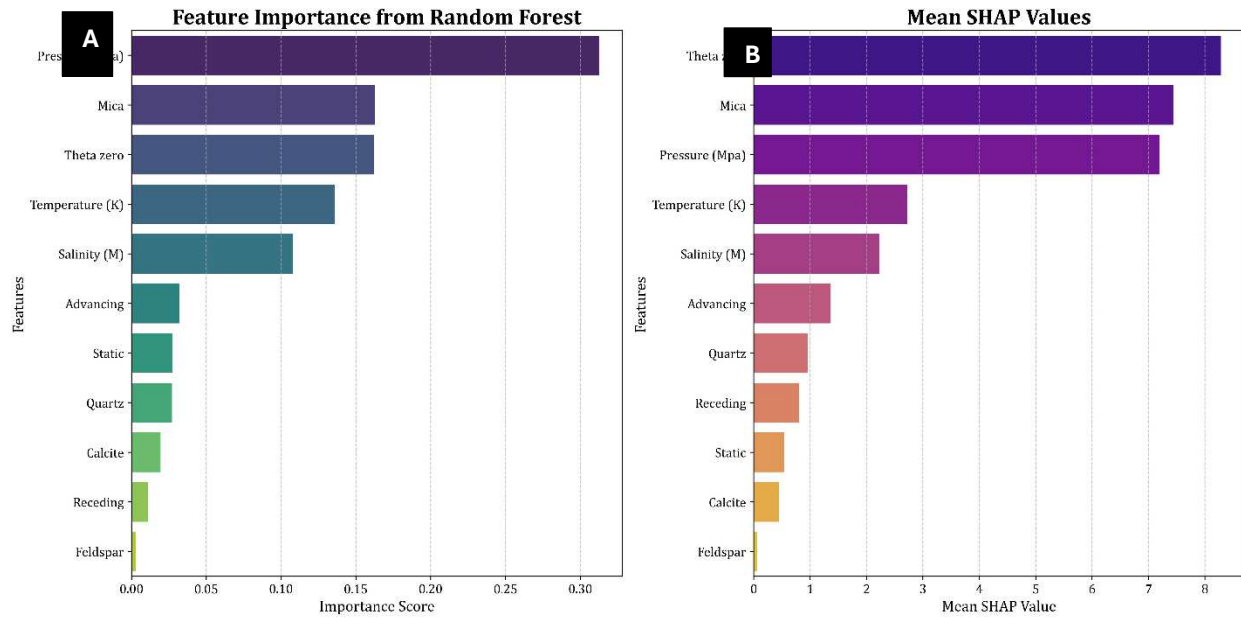


Figure 12. Comprehensive feature importance analysis and model interpretability visualization for the predicted  $\text{CO}_2$  contact angles, contrasting the variance-based importance scores derived from RF algorithm with the game-theoretic mean SHAP values to quantify the hierarchical impact of mineralogical and thermodynamic variables

It is decisively established by the synthesis of these dual interpretability metrics that precise characterization of reservoir pressure and initial baseline wetting states is absolutely mandatory for the reliable mathematical simulation of  $\text{CO}_2$  geo-sequestration operations. Furthermore, it is indicated by the pronounced algorithmic sensitivity to the specific inclusion of Mica, as opposed to inert minerals like Quartz or Calcite, that the heterogeneous mineralogical composition of the target saline aquifer will disproportionately govern capillary breakthrough pressures and ultimate residual trapping capacities. By mathematically isolating the principal thermodynamic drivers and demonstrating the diminished significance of intermediate advancing or receding measurements in global predictive frameworks, the required experimental parameter space is effectively streamlined by the deployed machine learning architecture. Ultimately, the optimization of site-selection criteria is directly facilitated by these advanced analytical insights, ensuring that continuous injection strategies are dynamically calibrated against the most sensitive operational

and geochemical constraints to maximize macroscopic storage security and mitigate the long-term risk of supercritical CO<sub>2</sub> leakage.

#### **4. Conclusion**

This study established a robust computational workflow to forecast the equilibrium contact angle of the carbon dioxide-brine-mineral interface, which is a strictly essential parameter for evaluating geological carbon sequestration viability. By compiling a comprehensive dataset of 630 empirical data points and employing a rigorous five-fold cross-validation framework, fifteen diverse machine learning algorithms were systematically evaluated to determine the optimal predictive matrix. The analytical findings definitively demonstrated the structural inadequacy of linear mathematical models in capturing highly non-linear thermodynamic behaviors and visually highlighted the severe overfitting tendencies of unregularized decision trees. The CatBoost algorithm unequivocally emerged as the premier predictive architecture, delivering unparalleled accuracy with a testing  $R^2$  of 0.9357 and exhibiting exceptional structural stability across extensive visual and statistical residual error diagnostics. Furthermore, SHAP-based interpretability analysis successfully isolated operational pressure, baseline wettability ( $\theta_0$ ), and the specific mineralogical presence of phyllosilicates like mica as the fundamental primary drivers of multiphase interfacial dynamics, effectively streamlining the required geochemical parameter space. This highly reliable predictive framework provides critical quantitative insights for optimizing structural and residual trapping mechanisms during subsurface carbon injection. However, the current predictive methodology is inherently constrained by the static nature of the aggregated tabular data, which may not completely capture the dynamic, long-term geochemical degradation and localized mineral dissolution occurring over extended geological timescales. Future research directions must fundamentally focus on integrating time-series experimental data and coupled reactive transport

modeling to continuously adapt the algorithmic framework to evolving subsurface mineralogical alterations.

### **Funding Statement**

No Funding.

### **Authors contribution statement**

Mohammadreza Kazemi: Data curation management, Modeling, Data curation, Drafted the initial manuscript, Reviewed and edited the manuscript.

### **Data Availability Statement**

The data that support the findings of this study are available from the corresponding author, A. Baghban, upon reasonable request.

### **References**

1. Soltanian, M.R., et al., *Multicomponent reactive transport of carbon dioxide in fluvial heterogeneous aquifers*. Journal of Natural Gas Science and Engineering, 2019. **65**: p. 212–223.
2. Bemani, A., et al., *Estimation of adsorption capacity of CO<sub>2</sub>, CH<sub>4</sub>, and their binary mixtures in Quidam shale using LSSVM: Application in CO<sub>2</sub> enhanced shale gas recovery and CO<sub>2</sub> storage*. Journal of Natural Gas Science and Engineering, 2020. **76**: p. 103204.
3. Zhang, Z. and D. Huisingh, *Carbon dioxide storage schemes: technology, assessment and deployment*. journal of cleaner production, 2017. **142**: p. 1055–1064.
4. Aghdam, S.K.-y., et al., *Thermodynamic Modeling of Saponin Adsorption Behavior on Sandstone Rocks: An Experimental Study*. Arabian Journal for Science and Engineering, 2023. **48**(7): p. 9461–9476.
5. Khezerlooe-ye Aghdam, S., et al., *Mechanistic assessment of Seidlitzia Rosmarinus-derived surfactant for restraining shale hydration: A comprehensive experimental investigation*. Chemical Engineering Research and Design, 2019. **147**: p. 570–578.
6. Chen, L., D. Liu, and R. Agarwal, *Numerical simulation of enhancement in CO<sub>2</sub> sequestration with various water production schemes under multiple well scenarios*. Journal of Cleaner Production, 2018. **184**: p. 12–20.
7. Jing, J., Y. Yang, and Z. Tang, *Effects of formation dip angle and salinity on the safety of CO<sub>2</sub> geological storage-a case study of Shiqianfeng strata with low porosity and low permeability in the Ordos Basin, China*. Journal of Cleaner Production, 2019. **226**: p. 874–891.

8. Hesse, M.A., F.M. Orr, and H.A. Tchelepi, *Gravity currents with residual trapping*. Journal of Fluid Mechanics, 2008. **611**: p. 35–60.
9. Juanes, R., et al., *Impact of relative permeability hysteresis on geological CO<sub>2</sub> storage*. Water resources research, 2006. **42**(12).
10. Iglaier, S., *Dissolution trapping of carbon dioxide in reservoir formation brine-a carbon storage mechanism*. 2011: INTECH Open Access Publisher London, UK.
11. Gaus, I., *Role and impact of CO<sub>2</sub>-rock interactions during CO<sub>2</sub> storage in sedimentary rocks*. International journal of greenhouse gas control, 2010. **4**(1): p. 73–89.
12. Al-Khdheawi, E.A., et al., *Influence of CO<sub>2</sub>-wettability on CO<sub>2</sub> migration and trapping capacity in deep saline aquifers*. Greenhouse Gases: Science and Technology, 2017. **7**(2): p. 328–338.
13. Iglaier, S., M.S. Mathew, and F. Bresme, *Molecular dynamics computations of brine-CO<sub>2</sub> interfacial tensions and brine-CO<sub>2</sub>-quartz contact angles and their effects on structural and residual trapping mechanisms in carbon geo-sequestration*. Journal of colloid and interface science, 2012. **386**(1): p. 405–414.
14. Pentland, C.H., et al., *Measurements of the capillary trapping of super-critical carbon dioxide in Berea sandstone*. Geophysical Research Letters, 2011. **38**(6).
15. Qi, R., T.C. LaForce, and M.J. Blunt, *Design of carbon dioxide storage in aquifers*. International Journal of Greenhouse Gas Control, 2009. **3**(2): p. 195–205.
16. Raza, A., et al., *Integrity analysis of CO<sub>2</sub> storage sites concerning geochemical-geomechanical interactions in saline aquifers*. Journal of Natural Gas Science and Engineering, 2016. **36**: p. 224–240.
17. Farokhpour, R., et al., *Wettability behaviour of CO<sub>2</sub> at storage conditions*. International Journal of Greenhouse Gas Control, 2013. **12**: p. 18–25.
18. Chen, C., et al., *Pressure and temperature dependence of contact angles for CO<sub>2</sub>/water/silica systems predicted by molecular dynamics simulations*. Energy & Fuels, 2016. **30**(6): p. 5027–5034.
19. Chen, C., et al., *Water contact angles on quartz surfaces under supercritical CO<sub>2</sub> sequestration conditions: Experimental and molecular dynamics simulation studies*. International Journal of Greenhouse Gas Control, 2015. **42**: p. 655–665.
20. Mutailipu, M., et al., *Measurement and estimation of CO<sub>2</sub>-brine interfacial tension and rock wettability under CO<sub>2</sub> sub-and super-critical conditions*. Journal of colloid and interface science, 2019. **534**: p. 605–617.
21. Voskarides, K., et al., *Prevalence of autoimmune diseases is strongly associated with average annual temperatures: systematic review and linear regression analysis*. BMC Rheumatology, 2025. **9**(1): p. 86.
22. Ezenarro, J. and D. Schorn-García, *How Are Chemometric Models Validated? A Systematic Review of Linear Regression Models for NIRS Data in Food Analysis*. Journal of Chemometrics, 2025. **39**(6): p. e70036.
23. Liu, R. and G. Yu, *Estimation of the Error Structure in Multivariate Response Linear Regression Models*. WIREs Computational Statistics, 2025. **17**(2): p. e70021.
24. Castillo, A., et al., *When Do Interaction/Moderation Effects Stabilize in Linear Regression?* Advances in Methods and Practices in Psychological Science, 2026. **9**(1): p. 25152459251407860.
25. Akhtar, N. and M.F. Alharthi, *Enhancing accuracy in modelling highly multicollinear data using alternative shrinkage parameters for ridge regression methods*. Scientific Reports, 2025. **15**(1): p. 10774.

26. Gillariose, J., J. Joseph, and C. Chesneau, *Lasso and Ridge regression: a comprehensive review of applications and developments in machine learning*. International Journal of Data Science and Analytics, 2025. **21**(1): p. 7.
27. Han, C., et al., *An optimized ridge regression for forecasting time series with a fixed period*. Pattern Recognition Letters, 2025. **197**: p. 274–281.
28. Bernardini, D., S. Paterlini, and E. Taufer, *New estimation approaches for graphical models with elastic net penalty*. Econometrics and Statistics, 2025. **33**: p. 258–281.
29. Nayem, H.M., S. Aziz, and B.M.G. Kibria *Evaluating Estimator Performance Under Multicollinearity: A Trade-Off Between MSE and Accuracy in Logistic, Lasso, Elastic Net, and Ridge Regression with Varying Penalty Parameters*. Stats, 2025. **8**, 45 DOI: 10.3390/stats8020045.
30. Yang, S., et al., *Enhanced soft Monte Carlo simulation coupled with support vector regression for structural reliability analysis*. Proceedings of the Institution of Civil Engineers - Transport, 2024. **178**(7): p. 459–474.
31. Xia, X., et al., *State of health estimation for lithium-ion batteries based on impedance feature selection and improved support vector regression*. Energy, 2025. **326**: p. 136135.
32. Sánchez, J.C.M., et al., *Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting*. Smart Agricultural Technology, 2025. **10**: p. 100791.
33. Qi, Z., et al., *Enhancing hydropower generation Predictions: A comprehensive study of XGBoost and Support Vector Regression models with advanced optimization techniques*. Ain Shams Engineering Journal, 2025. **16**(1): p. 103206.
34. Pantic, I.V., et al., *Artificial intelligence – based approaches based on random forest algorithm for signal analysis: Potential applications in detection of chemico - biological interactions*. Chemico-Biological Interactions, 2025. **418**: p. 111624.
35. Yang, Y. and H. Wang *Random Forest-Based Machine Failure Prediction: A Performance Comparison*. Applied Sciences, 2025. **15**, 8841 DOI: 10.3390/app15168841.
36. Sabuj Kumar, S., *AI DRIVEN PREDICTIVE MAINTENANCE IN PETROLEUM AND POWER SYSTEMS USING RANDOM FOREST REGRESSION MODEL FOR RELIABILITY ENGINEERING FRAMEWORK*. American Journal of Scholarly Research and Innovation, 2025. **4**(01): p. 363–391.
37. Rizkallah, L.W., *Enhancing the performance of gradient boosting trees on regression problems*. Journal of Big Data, 2025. **12**(1): p. 35.
38. Le Nguyen, K., M. Shakouri, and L.S. Ho, *Investigating the effectiveness of hybrid gradient boosting models and optimization algorithms for concrete strength prediction*. Engineering Applications of Artificial Intelligence, 2025. **149**: p. 110568.
39. Chen, Z., et al., *Flow mechanism of a vibrating prism using the combined K-nearest neighbor and dynamic mode decomposition method*. Journal of Wind Engineering and Industrial Aerodynamics, 2025. **258**: p. 106013.
40. Amer, A.A., S.D. Ravana, and R.A.A. Habeeb, *Effective k-nearest neighbor models for data classification enhancement*. Journal of Big Data, 2025. **12**(1): p. 86.
41. Hu, L., et al., *Significance-based decision tree for interpretable categorical data clustering*. Information Sciences, 2025. **690**: p. 121588.
42. Lim-Apo, F.A., F. Oliveira, and S. Hamacher, *Decision tree enhancer (DTE): Improving decision trees with optimization*. Knowledge-Based Systems, 2026. **339**: p. 115597.
43. Mienye, I.D., et al. *Deep Convolutional Neural Networks in Medical Image Analysis: A Review*. Information, 2025. **16**, 195 DOI: 10.3390/info16030195.

44. Raj, R. and A. Kos *An Extensive Study of Convolutional Neural Networks: Applications in Computer Vision for Improved Robotics Perceptions*. Sensors, 2025. **25**, 1033 DOI: 10.3390/s25041033.
45. Ding, H., et al. *Application of Convolutional Neural Networks and Recurrent Neural Networks in Food Safety*. Foods, 2025. **14**, 247 DOI: 10.3390/foods14020247.
46. Daryasafar, A. and K. Shahbazi, *Prediction of CO<sub>2</sub>/water/quartz wettability behavior during CO<sub>2</sub> storage in deep saline aquifers*. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2023. **45**(3): p. 9593–9601.
47. Daryasafar, A., A. Keykhosravi, and K. Shahbazi, *Modeling CO<sub>2</sub> wettability behavior at the interface of brine/CO<sub>2</sub>/mineral: Application to CO<sub>2</sub> geo-sequestration*. Journal of Cleaner Production, 2019. **239**: p. 118101.
48. Teodorescu, V. and L. Obreja Braşoveanu *Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost*. Computation, 2025. **13**, 127 DOI: 10.3390/computation13050127.
49. Gorriz, J.M., et al., *Is K-fold cross validation the best model selection method for machine learning?* Information Fusion, 2026. **135**: p. 104404.
50. Wang, Y., et al., *Outlier Identification Method Based on Multi-Model Weighted Consensus in Conjunction With Monte Carlo Cross-Validation*. Journal of AOAC INTERNATIONAL, 2025. **108**(6): p. 818–827.
51. Ergin, M., Y. Aksu, and Ö. Koşkan, *A comprehensive benchmarking of different outlier detection techniques for univariate datasets by Monte Carlo simulations*. Communications in Statistics - Simulation and Computation, 2025: p. 1–16.