Research article

Investigation of wettability and IFT alteration during hydrogen storage using machine learning

Mehdi Maleki [a], Mohammad Rasool Dehghani [a], Ali Akbari [a], Yousef Kazemzadeh [a,b,*], Ali Ranjbar [a]

[a] Department of Petroleum Engineering, Faculty of Petroleum, Gas, and Petrochemical Engineering, Persian Gulf University, Bushehr, Iran
[b] Persian Gulf University-Northeast Petroleum University of China Joint Research Laboratory, Oil and Gas Research Center, Persian Gulf University, Bushehr, Iran

ABSTRACT

Reducing the environmental impact caused by the production or use of carbon dioxide (CO_2) and other greenhouse gases (GHG) has recently attracted the attention of scientific, research, and industrial communities. In this context, oil production and enhanced oil recovery (EOR) have also focused on using environmentally friendly methods. CO_2 has been studied as a significant gas in reducing harmful environmental effects and preventing its release into the atmosphere. This gas, along with methane (CH_4) and nitrogen (N_2), is recognized as a 'cushion gas'. Given that hydrogen (H_2) is considered a green and environmentally friendly gas, its storage for altering wettability (contact angle (CA) and interfacial tension (IFT)) has recently become an intriguing topic. This study examines how H_2 can be utilized as a novel cushion gas in EOR systems. In this research, the role of H_2 and its storage in altering wettability in the presence of other cushion gases has been investigated. The performance of H_2 in changing the CA and IFT with other gases has also been compared using machine learning (ML) models. During this process, ML and experimental data were used to predict and report the values of IFT and CA. The data used underwent statistical and quantitative preprocessing, processing, evaluation, and validation, with outliers and skewed data removed. Subsequently, ML models such as Random Forest (RF), Random Tree, and LSBoost were implemented on training and testing data. During this process of modeling and predicting IFT and CA, the hyperparameters were optimized using Bayesian algorithms and random search (RS) methods. Finally, the results and performance of the modeling were evaluated, with the LSBoost modeling method using Bayesian optimization reporting R^2 values of 0.998614 for IFT and 0.986999 for CA.

1. Introduction

Carbon dioxide (CO_2) is recognized as a significant greenhouse gas (GHG) and is responsible for over 50 % of the overall environmental damage [1–3]. In recent decades, efforts to improve quality of life have led to a substantial reduction in the exploitation of

---

natural resources, largely due to the recognition of fossil fuel consumption—as a primary contributor to GHG emissions—as a major environmental threat. Consequently, there has been an increased focus on lowering GHG levels and exploring alternative sources that produce fewer or no harmful emissions [4–10]. Additionally, energy systems with minimal reliance on hydrocarbons have become a focal point. This shift underscores the growing need for renewable energy sources, which is closely linked to advancements in large-scale storage technologies. Large-scale energy storage, particularly for hydrogen (H_2), is essential as it enables the reliable integration of renewable energy into the global energy system. H_2, when converted into sustainable molecules, can be stored extensively. These sustainable molecules can be injected and stored in underground formations, such as hydrocarbon reservoirs [11–14]. Therefore, developing large-scale energy storage technologies is crucial for the transition to renewable energy systems [3].

Energy must be transformed into forms that facilitate large-scale storage. H_2 is one of the most prominent energy carriers, noted for its high energy density and the absence of carbon emissions when used as a fuel. Recently, there has been a significant global push to mitigate CO_2 and GHG emissions through the adoption of H_2 [15–17]. In this context, H_2 is increasingly being considered as a replacement for fossil fuels. While renewable energy sources like wind and solar power offer carbon-free alternatives, their effectiveness is constrained by seasonal variability and they are not always reliable for consistent energy supply. H_2 presents a promising option for the evolution of global energy systems. It can be utilized across various sectors, including power generation, food production, agriculture, metal industries, petrochemicals, electronics, and transportation, due to its environmentally friendly properties [18–20].

In the H_2 economy value chain, underground hydrogen storage (UHS) is essential [21–24]. Due to H_2's low volumetric energy density, it must be stored underground at high pressures to ensure sufficient availability for extraction and supply as needed [25,26]. Several studies have highlighted various challenges associated with storing H_2 in geological formations, such as salt caverns, aquifers, and depleted oil and gas reservoirs [27–30]. Storage challenges include H_2 leakage, chemical interactions, and biofilm formation. Basalt formations, deep aquifers, and depleted hydrocarbon reservoirs are ideal for storage due to their high capacity and cap rock properties. Depleted natural gas reservoirs are also favorable for H_2 and CO_2 storage. H_2 and natural gas storage differ in properties and processes. A lack of specific characterization data for accurate reservoir simulation is a challenge for UHS [31–34].

Experience with H_2 storage in porous reservoirs is limited, with ongoing projects exploring this area. Storage sites are chosen based on total capacity, including both working gas (recoverable H_2) and cushion gas (to maintain pressure). Non-H_2 gases like CH_4, CO_2, and N_2 are preferred for cushion gas due to lower cost and favorable properties. CO_2 is often chosen for its GHG emission reduction benefits. The ratio of cushion gas to H_2 varies based on reservoir conditions.

Effective UHS requires a thorough understanding of capillary pressure, which is crucial for preventing H_2 from migrating towards the cap rock (structural trapping). Capillary pressure determines the capillary pressure coefficient and the efficiency of a cap rock, indicating the point at which H_2 gas, being non-wetting, enters the largest pores of the cap rock at the storage site. If the storage site has been previously saturated with a wetting phase, such as water or brine, this could lead to the displacement of H_2. The Laplace-Young equation (Equation (1)) is used to estimate the capillary pressure (or entry capillary pressure, Pce) for a cap rock characterized by cylindrical-shaped pores with a maximum radius, r [35].

$$P_{ce} = P_{H_2} - P_{H_2O/brine} = \frac{2\gamma_{H_2O/brine}\ Cos\ \theta}{r} \tag{1}$$

in this equation, θ represents the contact angle (CA) of the H_2/brine/rock system, and γ denotes the surface tension (ST) or IFT between the H_2 and H_2O/brine phases. It is important to note that effective H_2 storage is achieved when the capillary entry pressure (P_ce) exceeds the buoyancy pressure exerted by the fluids in the cap rock and reservoir [36].

Capillary forces, driven by interfacial tension (IFT) and wettability, play a crucial role in fluid displacement within porous media. Wettability, which reflects the affinity of the solid surface for a fluid, impacts how effectively fluids can move through porous rock formations. In capillary-dominated systems, such as in fluid displacement in porous rocks, lower IFT facilitates the displacement of wetting fluids in imbibition processes, while resisting forces act in drainage when the non-wetting phase displaces the wetting phase. Nanofluids can further influence capillary forces by modifying IFT and wettability. For example, nanofluid-assisted displacement reduces IFT, creating structural disjoining pressure, which helps spread the fluid uniformly at the pore scale. This leads to more efficient fluid movement during imbibition, as lower IFT increases favorable capillary forces, improving wetting-phase invasion. Similarly, during drainage, nanofluids reduce the resistance from capillary forces, enhancing recovery efficiency [37]. In systems using interfacially active nanoparticles and low-salinity brines, nanoparticles reduce IFT and shift wettability toward more water-wet conditions. This enhances imbibition, as the favorable capillary forces drive wetting fluids into the pores. In drainage processes, the combination of nanoparticles and low-salinity brines weakens capillary resistance, facilitating more efficient fluid displacement and improving recovery outcomes. Understanding and controlling these forces is essential for optimizing fluid displacement in porous media [38].

Extensive research has been conducted on H_2 storage in porous media, including both direct quantitative studies of wettability and indirect qualitative assessments using core flooding and NMR to measure dynamic and static CAs. Studies have examined the interfacial properties between gas/liquid (H_2/brine) and rock/liquid (rock/H_2 and rock/brine). IFT and CA are critical factors in underground H_2 storage, as changes in these properties can influence the adsorption, distribution, and retention of H_2 within reservoirs. A thorough understanding of these properties is crucial for enhancing H_2 storage efficiency. Accurate measurement of these properties under varying conditions is essential, and numerical methods are often employed alongside experimental approaches. However, these measurements can be labor-intensive and expensive, especially at high temperatures and pressures.

Recent research by Hosseini and Leonenko [39] has introduced three empirical models for estimating the IFT between pure H_2 and

brine under various subsurface storage conditions. These models account for factors such as temperature, pressure, and salt salinity, providing accurate predictions of IFT within specified temperature and pressure ranges. Additionally, machine learning (ML) techniques have emerged as valuable tools for optimizing H_2 storage. Thanh and colleagues explored H_2 compressibility using four ML algorithms—Random Forest, Light Gradient Boosting, Extreme Gradient Boosting, and AdaBoost_Decision Tree. Among these, XGBoost demonstrated the highest performance, achieving an R^2 value of 0.941. Another study developed four artificial intelligence models to forecast the IFT between H_2 and salt based on pressure, temperature, and molality. The Random Forest (RF) model showed notable superiority, with high accuracy and R^2, RMSE, and AARD values of 0.96, 1.50, and 1.84 percent, respectively.

Ng et al. [40] used ML methods to predict IFT in the H2-brine system, testing four techniques: GBR, GP, MLP with LMA, and MLP with Adam. The MLP-LMA model was the most effective, with an R^2 value of 0.9997. Behnamnia et al. [41] developed AI models to predict IFT between water and various gases (H_2, H_2 + edge gas, CH_4, CO_2, N_2) and reported satisfactory performance.

In N. Suetrong et al. [42] study, the researchers aim to develop a machine learning model predicting dynamic contact angles during immiscible fluid displacement in subsurface environments. Using a dataset of experimental measurements, the study focuses on capturing temporal variations of contact angles critical for understanding fluid behavior in porous media. Results validate the effectiveness of their approach in accurately predicting dynamic contact angles, enhancing predictive capabilities for subsurface flow applications. This research contributes to advancing knowledge of fluid-solid interactions under dynamic conditions, promising improved strategies for hydrocarbon recovery and environmental remediation.

Thanh et al. [43]. review the application of data-driven machine learning models to predict hydrogen solubility in varying salinity aqueous systems, focusing on implications for underground hydrogen storage. They utilize Gaussian Process Regression (GPR), Least Squares Support Vector Machine (LSSVM), and Extra Tree (ET) methods, achieving high accuracy with R^2 values close to 1 and low RMSE and MAE. This research enhances understanding of optimal hydrogen storage in saline aquifers, offering insights into effective strategies for future applications.

Masoudi et al. (2024) [44] investigate the potential of lined rock caverns as an innovative approach for hydrogen storage. Utilizing a blend of numerical simulations and practical field tests, the researchers assess the caverns' structural durability, storage capacity, and operational safety. Their findings reveal that lined rock caverns are capable of storing substantial quantities of hydrogen while ensuring structural stability across a range of pressure conditions. The study underscores the benefits of this storage method, including superior safety and efficiency compared to traditional solutions. The results indicate that lined rock caverns represent a viable and advanced option for large-scale hydrogen storage, effectively addressing critical safety and capacity concerns.

Hassanpouryouzband et al. (2024) [45] examine the future prospects of hydrogen energy, contrasting the "foraging" approach with the "farming" strategy. Their study integrates a detailed review of current hydrogen production technologies with scenario modeling to forecast future developments and their implications. The findings indicate that while the foraging approach, which relies on existing and often less sustainable methods, offers short-term solutions, the farming strategy—focused on advancing technology and sustainable practices—presents a more viable and long-term solution. The research highlights the advantages of adopting hydrogen farming to improve efficiency, lower costs, and facilitate a broader energy transition. This analysis emphasizes the importance of transitioning towards farming methods to ensure a sustainable and economically viable future for hydrogen energy.

In addition to the insights provided by previous studies, we have incorporated a comprehensive table into the manuscript to further elucidate the application of machine learning techniques in similar research areas. Table 1 presents a detailed literature review, summarizing key works that utilize machine learning to explore topics related to flow in porous media and hydrogen storage. By including this table, we aim to enhance the depth of our literature review and provide a clearer context for our study, illustrating how machine learning methods have been applied across various relevant domains.

This paper explores the role of ML in analyzing large and complex laboratory data sets. Unlike traditional methods, which can be time-consuming and expensive, ML offers significant advantages in data optimization and analysis. It improves the ability to predict future outcomes, uncover patterns, and reveal hidden relationships, leading to enhanced experimental processes, reduced material usage, and increased efficiency. In this study, ML algorithms were used to predict, assess performance, and optimize data related to CAs and IFT of H_2 in the presence of brine and protective gases. This data was derived from previous research on UHS.

The utilization of cushion gas in hydrogen storage systems marks a significant advancement in optimizing storage efficiency and ensuring operational stability. In this article, the application of cushion gas has been explored with a focus on its role in maintaining reservoir pressure, thereby mitigating pressure fluctuations and enhancing the overall stability of hydrogen storage systems. Empirical data and field observations incorporated in this study underscore the efficacy of cushion gas in preventing hydrogen leakage and preserving reservoir integrity, contributing to a more reliable and secure storage solution. This approach not only stabilizes pressure levels but also fortifies the operational stability of hydrogen storage systems, addressing key safety and performance concerns. Moreover, this article leverages advanced experimental and real-world data to elucidate the substantial economic advantages and improved reservoir management practices associated with cushion gas. The integration of cushion gas, as analyzed in the study, reveals significant reductions in operational costs and enhances the economic viability of hydrogen storage systems. By employing sophisticated models, including those enhanced with machine learning techniques, the study provides nuanced insights into the dynamic interactions between hydrogen, cushion gas, and reservoir rock. These findings lead to more accurate performance predictions and optimized design strategies, advancing the long-term sustainability and economic feasibility of large-scale hydrogen storage initiatives.

2. Data collection and specific descriptions

In the current study, data on pressure, temperature, salinity, and various compositions of protective gases (CO_2, N_2, CH_4) were used

Table 1
Literature review (hydrogen storage using machine learning).

author	Year of publication	Methodology	Machine learning models	The best final result	Results of the article
C.S.W. Ng et al [46]	2022	Modeling interfacial stress in saline water-hydrogen systems using advanced machine learning techniques.	Gradient Boosting Regressor (GBR), Genetic Programming (GP), MLP-LMA, MLP-Adam	MLP-LMA with $R^2 = 0.9997$	The study investigates modeling interfacial stress between hydrogen and saline water using advanced machine learning techniques, showing MLP-LMA performs the best.
H.V. Thanh et al. [47]	2023	Predicting corrosion of rock/mineral-water-saline water-hydrogen systems using multiple machine learning approaches.	Random Forest (RF), Light Gradient Boosting (LGRB), Extreme Gradient Boosting (XGBoost), Adaboost_DT	XGBoost with $R^2 = 0.941$	The research evaluates the impact of nanocomposites on rock corrosion and predicts outcomes using machine learning algorithms, where XGBoost achieves the highest accuracy.
Ahmad and Sigel [48]	2021	Use of laboratory data and simulations to predict hydrogen storage performance in MOFs.	Random Forest, Support Vector Machine (SVM), Neural Networks	Neural Networks with $R^2 = 0.92$	The paper demonstrates that machine learning models, especially neural networks, can provide more accurate predictions of hydrogen storage performance in MOFs, aiding in the optimization of MOF designs.
Mao et al. [49]	2024	Reservoir simulation using machine learning techniques and simulated data.	Random Forest, Gradient Boosting, Support Vector Machine (SVM)	Best result: Random Forest with $R^2 = 0.95$	Machine learning models accurately predicted hydrogen recovery yield and gas purity; methods had higher accuracy and lower computational cost compared to traditional simulations.
S. Mao, B. Chen, M. Morales, M. Malki et al. [50]	2024	Development of ROM model based on reservoir simulation data.	Deep Learning	Specific details on best numerical regression result not provided.	The paper analyzes the effect of gas shielding on hydrogen storage in porous rocks using reservoir simulation and deep learning, highlighting the importance of these effects in optimizing hydrogen storage.
Mahdi Kanaani et al. [51]	2023	Use of machine learning methods and optimization for hydrogen and carbon dioxide storage.	Multi-layer Perceptron (MLP)	MLP with $R^2 = 0.9988$	The selected model successfully predicted the accuracy of all three objective functions (NPV, hydrogen recovery factor, and CCS).
Helland et al. [52]	2023	Use of combined reservoir simulations to determine optimal production times (until well flow reaches 1 % mole fraction CO2) in ideal CO2-filled reservoirs during methane (CH4) injection and production cycles. Generated data used for developing Artificial Neural Networks (ANNs) to describe relationships between input parameters and optimal production times.	Artificial Neural Networks (ANNs)	RMSE <0.02, $R^2 > 0.99$	Reservoir height and permeability have the greatest impact on the time to achieve 1 % mole fraction CO2. ANNs identified as useful tools for establishing relationships between critical parameters in UNGS operations aiming for increased CH4 production time and higher gas delivery with minimal CO2 production.
Bakhshian et al. [53]	2022	Use of combined reservoir simulations to assess the impact of residual methane on CO2 storage efficiency in a representative reservoir model of "HC sand" in the High Island 24L field. Utilization of Artificial Neural Networks (ANNs) models for rapid CO2 storage modeling and Monte Carlo simulations for quantifying uncertainties of	Artificial Neural Networks (ANNs)	Developed ANN model for predicting CO2 storage efficiency and assessing uncertainty using Monte Carlo simulations.	Residual hydrocarbon gas negatively affects remaining trapping and CO2 dissolution, increasing pressure management and injectivity by enhancing CO2 mass mobility. The developed data-driven model serves as a powerful tool for screening wet areas of gas-depleted reservoirs to evaluate CO2 storage efficiency.

Table 1 (*continued*)

author	Year of publication	Methodology	Machine learning models	The best final result	Results of the article
		geological and reservoir parameters.			
HV Thanh et al. [54]	2022	Development of intelligent deliverability models for natural gas storage underground using machine learning techniques. Modeling based on data from 387 UNGS sites in depleted fields, aquifers, and salt domes.	Gaussian Process Regression (GPR), Least Squares Support Vector Machine (LSSVM), Extra Tree (ET)	For depleted fields: R^2 = 0.999999998, RMSE = 4.75E-06, MAE = 0.00021. For salt domes: R^2 = 0.987, RMSE = 0.0046, MAE = 0.11. For aquifers: R^2 = 0.999999997, RMSE = 7.1094E-06, MAE = 0.0002102.	GPR model outperforms LSSVM and ET. These models are proposed as fast and accurate templates for predicting UNGS deliverability in depleted fields, aquifers, and salt domes. Results contribute to a deeper understanding of machine learning's role in natural gas storage prediction.
Wang et al. [55]	2024	Analysis and prediction for carbon capture and storage (CCS) projects using machine learning models and deep learning. Hyperparameter optimization using simulated annealing algorithm and grid search strategy. Use of well log data to predict reservoir volume and CO_2 saturation.	Ridge Regression (RR), Random Forest (RF), Gradient Boosting Regression (GBR), Support Vector Regression (SVR), Artificial Neural Network (ANN)	For predicting reservoir volume and porosity: GBR and RF were superior models. For estimating CO_2 saturation: RF performed the best, followed by SVR, ANN, GBR, and RR.	GBR and RF were top models for predicting reservoir volume and porosity. RF showed the best performance for estimating CO_2 saturation. Key features included sigma and well pressure. Results of this study help fill the knowledge gap in carbon storage and meet the increasing demand for sustainable and cost-effective energy solutions.
Ali et al. [56]	2024	Use of advanced machine learning models to predict effusion characteristics of mineral/H2/saline water systems under geological storage conditions. Evaluation of effusion behavior under different pressure, temperature, and salinity conditions.	Fully Connected Neural Networks, Adaptive Gradient Boosting, Random Forests, Decision Trees, Extreme Gradient Boosting	RMSE: 0.214 (training), 0.810 (testing)	Machine learning models demonstrated high capability in predicting effusion characteristics of mineral/H2/saline water systems. Pressure had the greatest impact on system effusion. These models can effectively predict hydrogen storage capacities and ensure security in large-scale projects.
Xie et al. [57]	2024	Use of molecular dynamics simulations to create a comprehensive database of IFT (interfacial tension) for (H2 + CO2)-saline water and develop a machine learning-based IFT equation. Model validation with experimental data and existing IFT equations.	Machine Learning (ML)	R^2 = 0.902 compared to 601 experimental data points	The developed machine learning-based IFT equation can be easily implemented and utilized in reservoir simulations and other UHS applications, providing good performance in predicting interfacial tension (H2 + CO2)-saline water.

to evaluate IFT and CA in the context of H_2. The methodology included dataset creation, parameter tuning for ML models, and performance evaluation. A total of 2,250 data points for IFT and 950 data points for CA were sourced from prior experimental research. The study investigated the effects of different protective gas mixtures and H_2 compositions on permeability and gas-liquid IFT. Data were organized and analyzed using Microsoft Excel. Linear regression analysis, which yielded an R^2 value of 0.8703, confirmed the reliability of the data extraction process. The final dataset provided insights into the effects of H_2's IFT in the presence of protective gases under varying conditions of pressure, temperature, and salinity, and was used for training and evaluating ML models.

$$IFT_{Mixture} = f(P.T.Salinity)$$

$$\sigma_{Mixture} = g(P.T.Salinity)$$

P and T denote pressure and temperature, respectively. The above functions illustrate the ability of the proposed ML models to predict IFT in H_2 and protective gas mixtures across various salinity levels, temperatures, and pressures. The necessary statistical information is detailed in Table 1.

In this study, contact angle measurements is a method to characterize wettability, leveraging data obtained under controlled laboratory conditions specific to hydrogen and cushion gas systems. These measurements were integral to our analysis of capillary forces and fluid displacement in porous media, aligning closely with the experimental framework of our research. However, it is important to acknowledge the inherent limitations of static contact angle measurements, which may not fully capture the dynamic

nature of fluid-solid interactions during processes such as imbibition and drainage. Variations in surface roughness, medium heterogeneity, and potential contaminants further underscore the complexity and potential variability of contact angle data. Despite these constraints, the static contact angle data provided a robust foundation for our study, offering valuable insights into wettability dynamics relevant to hydrogen storage applications with cushion gas.

The collected data underwent statistical analysis following a review, with IFT and CA plotted separately against variables such as salinity, temperature, pressure, and the percentages of H_2, CO_2, N_2, and CH_4. The analysis included two box plots, two pair plots, and two heat maps.

Figs. 1 and 2 present box plots depicting the distribution, central tendency, and variability of seven variables related to IFT and CA. These plots display medians, interquartile ranges (IQRs), whiskers, and outliers. In the IFT plot, H_2 and pressure exhibit high medians of 54.00 % and 1750.00 psi, respectively, with broad IQRs indicating significant variability. Conversely, salinity, N_2, and CO_2 show lower medians and more confined IQRs. N_2 has first and third quartiles of 10.00 %, reflecting a very tight range, whereas CH_4 has a median of 21.33 %, and temperature has a median of 50.00 °C. For IFT, a median of 59.183 with a moderate IQR and several extreme values is observed. The pair plots in Fig. 3 explore the relationships between pairs of variables and the distribution of individual variables. Each diagonal plot features histograms or kernel density plots to show variable distributions, while scatter plots examine the
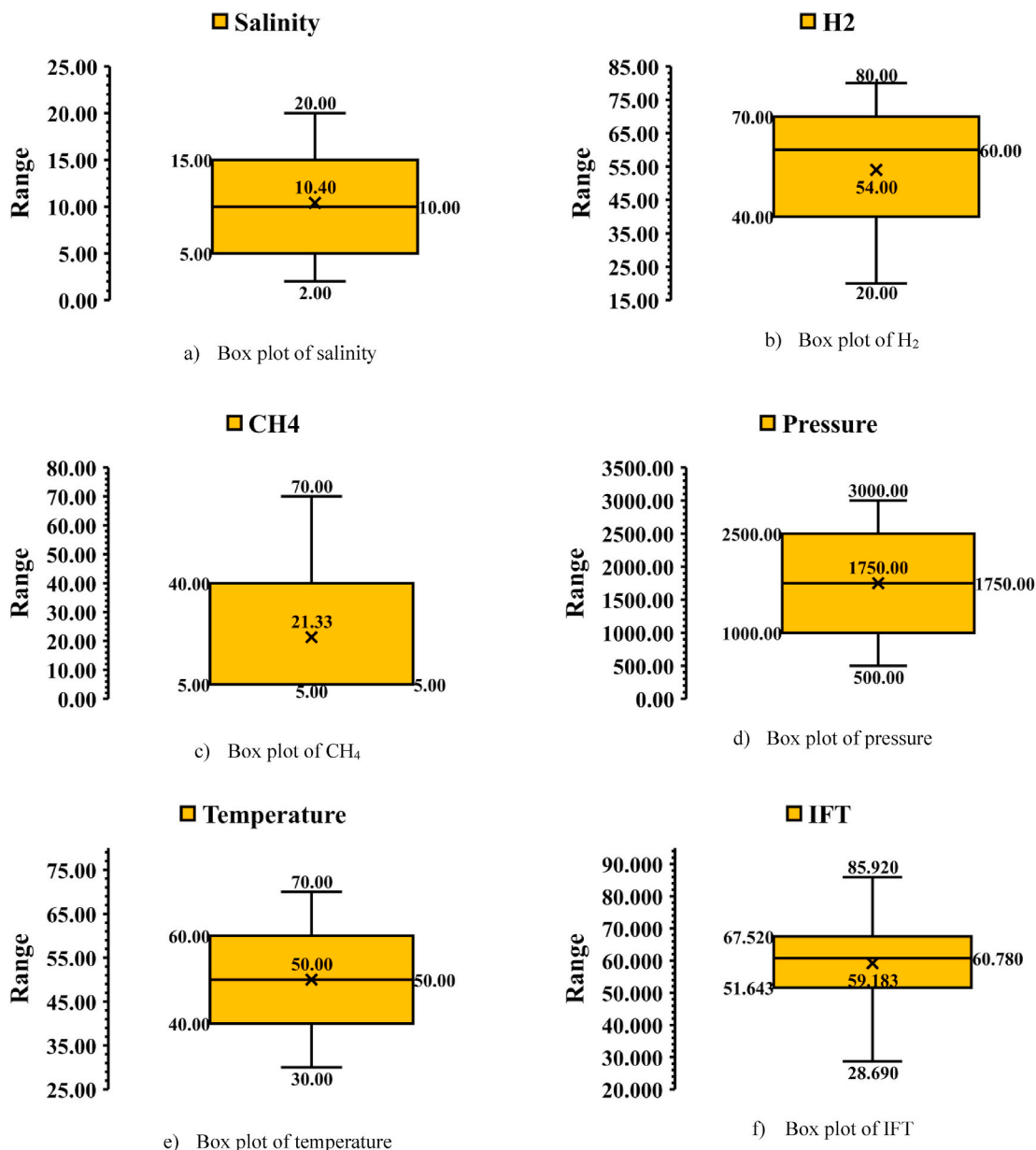

Fig. 1. Box-plot of IFT parameters: a) salinity, b) H_2, c) CH_4, d) Pressure, e) Temperature, f) IFT.
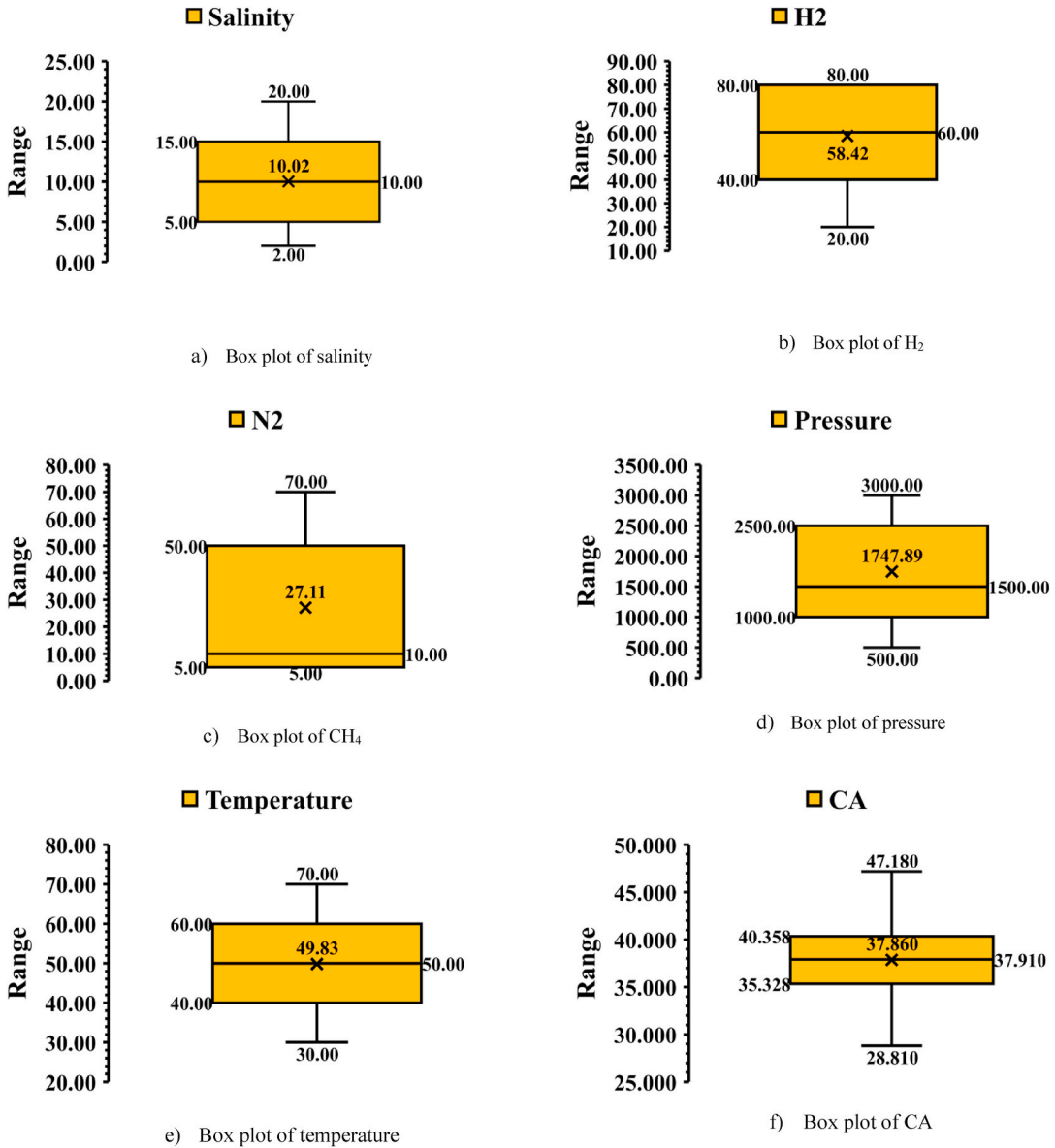
Fig. 2. Box-plot of CA's parameters: a) salinity, b) H$_2$, c) CH$_4$, d) Pressure, e) Temperature, f) CA.

relationships between variables.

Scatter plots help visually detect correlations and patterns between pairs of variables, while histograms and density plots on the diagonal display the distribution of individual variables. These plots also facilitate the identification of outliers that may require further scrutiny. Insights from these plots guide feature selection and engineering for ML models, complementing numerical correlation matrices by providing a visual representation of the strength and nature of relationships between variables.

Numerical correlation matrices are essential tools in ML and data analysis, illustrating linear relationships between variables. They play a critical role in feature selection, dimensionality reduction, and exploratory data analysis (EDA). In this study, a numerical correlation heatmap was employed, utilizing Pearson's correlation coefficient.

The Pearson correlation coefficient is a statistical metric that assesses the strength and direction of a linear relationship between two variables, with values ranging from −1 to 1. The possible values of this coefficient are:

- **Perfect Positive Correlation:** Indicates a simultaneous increase in the values of both variables.
- **Perfect Negative Correlation:** Indicates an inverse decrease in the values of both variables.
- **No Correlation:** Indicates the absence of a linear relationship between the two variables.
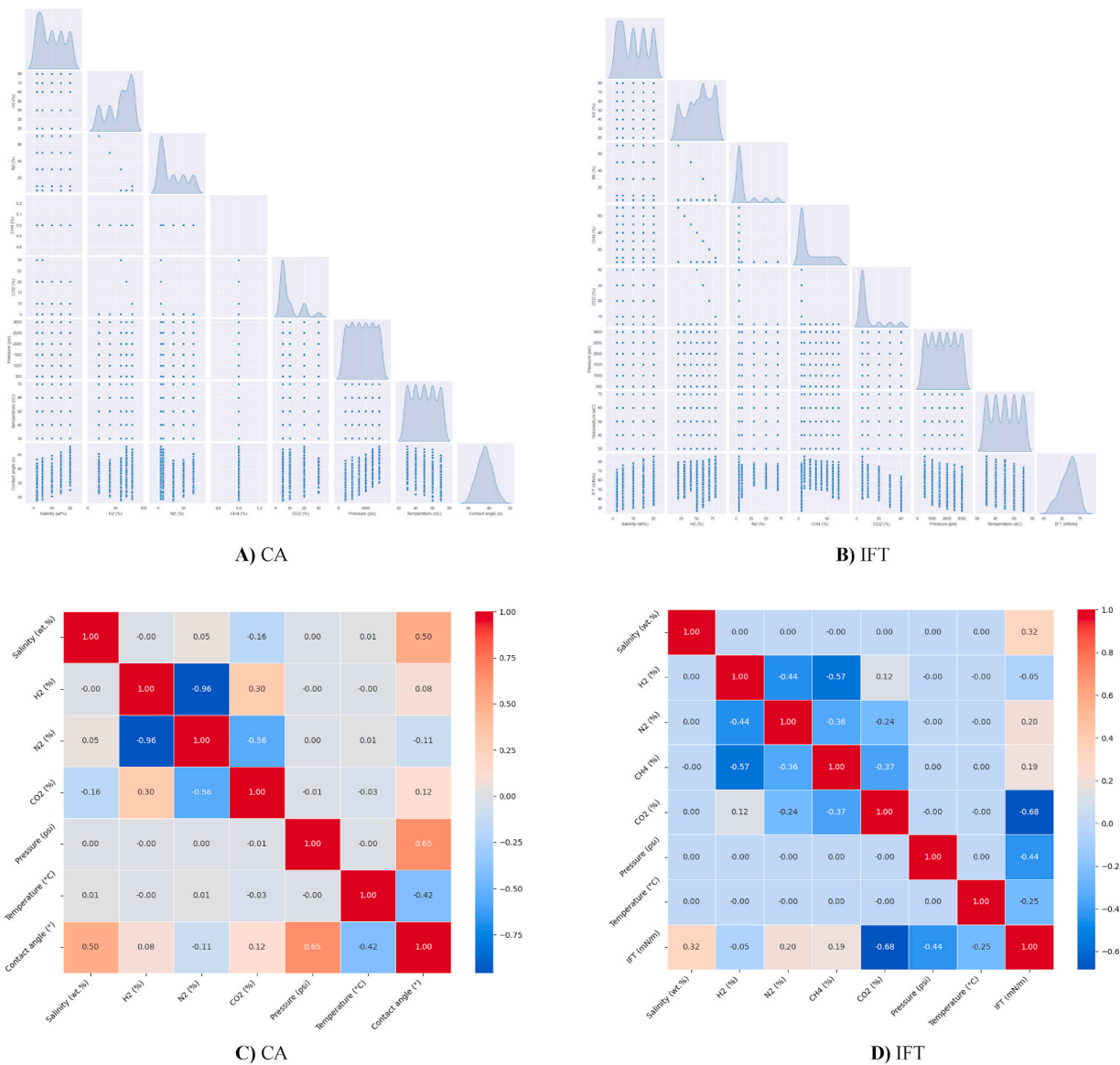
A) CA

B) IFT

C) CA

D) IFT

Fig. 3. Statistic plots and correlation diagram. A: CA pair plot; B: IFT pair plot; C: CA heat plot; D: IFT heat plot.

Equation (2) for the Pearson correlation coefficient is as follows:

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(X_i - \overline{X})^2 - (Y_i - \overline{Y})^2}}$$

2

in this equation, X_i and Y_i represent the observed values, while $\overline{X}$ and $\overline{Y}$ are the means of variables X and Y. Therefore, if $r > 0$, it indicates a positive (direct) correlation; the closer r is to 1, the stronger the positive relationship. Conversely, if $r < 0$, it signifies a negative (inverse) correlation; the closer r is to -1, the stronger the negative relationship. If $r = 0$, there is no linear relationship, and a non-linear relationship may exist (indicating no correlation).

In this study, the relationships among variables such as pressure, temperature, gas compositions, IFT, and CA were analyzed using the Pearson correlation coefficient and visualized with a heatmap. Fig. 3 illustrates the Pearson correlation coefficients between these variables, with a color gradient ranging from -0.6850 (yellow) to 1.000 (dark purple) for IFT, and from -0.9600 (yellow) to 1.000 (dark purple) for CA. The heatmap shows that IFT has correlations of -0.68 with CO_2, -0.44 with pressure, and -0.25 with temperature, while it has a positive correlation of 0.32 with salinity. For CA, the correlations are 0.65 with pressure and 0.50 with salinity, but -0.42 with temperature. This heatmap is valuable for optimizing processes and making informed decisions by providing a clear visual representation of the relationships between critical parameters in oil engineering.

3. Data processing

In this study, data extracted from three review articles and other relevant papers were analyzed [58]. These studies employed methods for measuring IFT and CA under laboratory conditions, with a particular focus on H_2. The research also examined the effects of protective gases such as H_2, N_2, CH_4, and CO_2 on H_2 storage capacity. Additionally, the evaluation and calculation of IFT and CA were influenced by parameters including pressure, temperature, and salinity.

The review studies provide various insights into the behavior of gas and water mixtures under different conditions of temperature, pressure, and salinity (Fig. 4). For instance, Mohammad et al. [58] (2023) found that in H_2 storage involving CH_4, the CA does not correlate with pressure (P), temperature (T), or salinity. However, the IFT was observed to decrease with higher temperature and pressure, while it increased with rising salinity. Similarly, Mohammad et al. [59] (2023) reported that the effect of N_2 on H_2 storage shows that the CA increases with pressure and decreases with temperature. In this study, IFT was found to decrease with increasing pressure and temperature but increase with salinity and the percentage of N_2.

IFT plays a critical role in determining the storage height within a reservoir by influencing the capillary forces in the porous medium, which affects the vertical extent of the stored H_2. Specifically, higher IFT values increase capillary forces, leading to a more pronounced capillary entry pressure at the interface between H_2 and the brine or other fluids present in the reservoir. This capillary entry pressure governs the maximum storage thickness that H_2 can achieve without breaching the cap rock seal. Conversely, lower IFT values may reduce capillary forces, potentially allowing H_2 to occupy a greater vertical extent within the reservoir.

A more detailed discussion on how variations in IFT impact the storage height of H_2 within different geological formations will be included in the manuscript. This addition will enhance the understanding of the interplay between IFT, capillary pressure, and storage capacity, which is crucial for optimizing UHS strategies.

3.1. Data preprocessing and sensitivity analysis

Data preprocessing is essential for preparing input data before it is used in models and algorithms. This process ensures that the data is accurate, correctly formatted, and suitable for further analysis. Key steps in data preprocessing include data cleaning (correcting errors and eliminating invalid entries), transforming data into suitable formats, normalizing to address scale differences, integrating data from various sources, handling missing values, and detecting noise. These steps can be carried out automatically or may require manual intervention and are critical for effective data analysis and ML in engineering contexts.

Data preprocessing encompasses several crucial steps to prepare data for ML models and data analysis. These steps include: data cleaning to eliminate incomplete or duplicate entries; converting data into standardized formats, such as transforming text into numerical vectors; feature selection and extraction to identify and use key attributes; dimensionality reduction to minimize complexity and reduce computational costs; data normalization to improve model performance; and noise removal to enhance data quality. Additionally, processing temporal data accurately is vital for models that depend on time-related information. These procedures ensure that data analysis models are accurate and effective.

Based on the points discussed, different data preprocessing techniques directly influence the improvement of modeling, prediction, and optimization outcomes for existing algorithms. Consequently, this study utilizes several data preprocessing methods, which will be described in detail below.

(i) While data preprocessing is crucial for effective data analysis across many applications, choosing inappropriate preprocessing methods can lead to the loss or distortion of valuable information that needs to be revealed during the analysis.
(ii) Effective data preprocessing requires a domain expert or thorough domain study for valuable insights and technique validation.
(iii) Data preprocessing is often iterative, with steps like data cleaning and feature selection repeated to achieve optimal results.

A key challenge in data preprocessing is to discern and preserve valuable information from raw data. The concept of data preprocessing can differ, with some views suggesting that it should incorporate feedback from data analysis. Research in this field focuses on creating tools that recommend suitable preprocessing techniques. With preliminary knowledge of the data and analysis goals, developing these tools becomes achievable.

(i) An expert system that analyzes data samples and user information to determine the best preprocessing strategy.
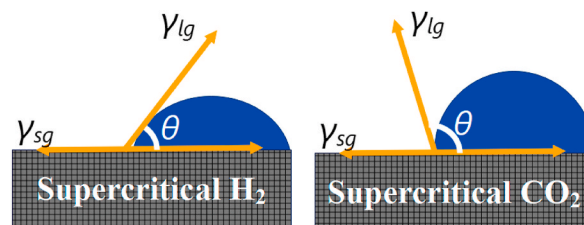

Fig. 4. Effect of gas mixture and cushion gas with water on IFT.

(ii) A toolbox with various data preprocessing techniques and practical examples, showing how different methods affect data analysis outcomes.

3.2. Cross-validation

In this study, during training model cross-validation method utilized to prevent overfitting on training data.

Cross-validation is a fundamental technique in machine learning and statistical modeling, widely utilized to evaluate the predictive performance of models while addressing challenges like overfitting and underfitting. The core concept involves dividing the dataset into distinct subsets, or folds, ensuring that each fold functions alternately as both a training and test set in successive iterations. The most common form, k-fold cross-validation, divides the dataset into k equally sized folds, with the model being trained on k-1 folds and validated on the remaining fold. This process is repeated k times, allowing each fold to serve as the test set once. This method provides a thorough assessment of the model's performance across different data subsets, enhancing its generalizability. Moreover, it aids in optimizing hyperparameters by yielding a more reliable estimation of the model's performance. Variants like stratified cross-validation, which maintain the distribution of target classes within each fold, are particularly useful when working with imbalanced datasets. By applying cross-validation, researchers gain a more accurate measure of a model's predictive capability on unseen data, rendering it an essential tool in the development of robust machine learning models [60].

3.3. Machine learning model

In this study, to predict IFT and CA, Regression Tree, Random Forest, and LSBoost were selected due to their effectiveness in handling datasets with a large number of variables and instances. Tree-based models are known for their simplicity, interpretability, and ability to capture non-linear relationships between features and target variables. Additionally, Random Forest and LSBoost, as ensemble methods, help improve model accuracy and reduce overfitting, making them well-suited for this predictive task. Besides these, Genetic expression programming was utilized to propose a correlation, which leads to a better understanding of the underlying relationships. Each of these methods is explained in detail below.

3.3.1. Genetic expression programming

Fig. 5 illustrates Genetic Expression Programming (GEP), which combines elements of Genetic Algorithms (GA) and GP. GEP is a robust method for tasks such as data mining, function discovery, optimization, and prediction. It evolves solutions to develop simple and accurate models without requiring prior knowledge of the model's structure or parameters. Its adaptability and effectiveness in tackling complex problems have led to its success across various fields. Similarly, GP, which integrates principles of genetic evaluation and neural networks, addresses complex issues through regression functions. GEP enhances genetic diversity and efficiently resolves intricate models by employing fixed chromosomes and tree-like structures.

During this study, after the collection and processing of data and the analysis of variable correlations, the data were fed into the Genetic Expression Programming (GEP) algorithm. To avoid the complexity of the developed relationships, only simple functions were employed in the model. The final form of the equation is presented in relationships 3 to 8. This equation consists of 9 constants, denoted as c_1 to c_9, which are listed in Table 2.

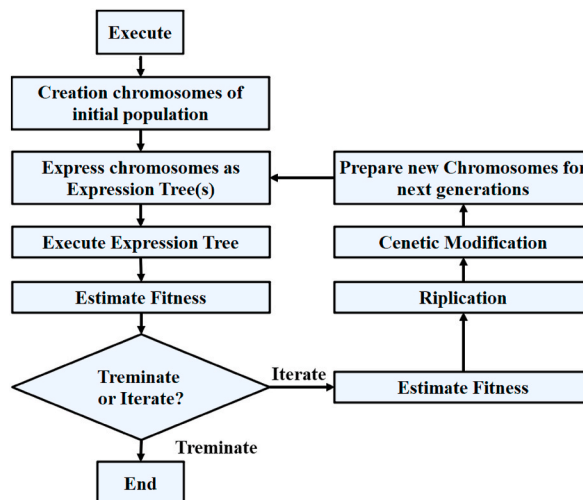$$N1 = \left(\frac{\ln(c1.T) + c2}{\ln 10} \right) \times \sqrt{H_2 - CO_2} \qquad\qquad 3$$

Fig. 5. Schematic diagram of GEP algorithm.

$$N2 = N1 + \ln\left(\left((c3 - N2).c4\right)^4 . \left((CO_2.H_2) + (Salinity)^5\right)\right) \tag{4}$$

$$N3 = N2 + \left(\frac{\sqrt{(c5.CO_2) + P + T}}{c6}\right) \tag{5}$$

$$N4 = N3 + + \left(10^{10^{c7}}\right)^{2.N_2.\ln(10) - 2.\ln(P)} \tag{6}$$

$$N5 = N4 + \left(\ln\left(\ln\left(10^{\left(\left(\frac{c8}{CO_2}\right)^{c9}\right)} + CH_4\right)\right)\right) \tag{7}$$

$$IFT = N5 + \left(\frac{H_2}{\left(\log_{10}(T)\right)^5 + H_2}\right).c10 \tag{8}$$

3.3.2. Decision tree

A prevalent method in data mining is the use of systems designed for classification tasks. Classification algorithms are effective at handling large datasets, aiding in the development of hypotheses for categorizing contextual classes, and classifying both training data and new information. This paper focuses on the DT algorithm within the realm of ML. As illustrated in Fig. 6, DTs are versatile tools used in diverse areas such as ML, image processing, and pattern recognition. These trees operate sequentially, integrating a series of base tests where each test evaluates a numerical feature against a specified threshold. Creating conceptual rules in DTs is generally more straightforward than establishing numerical weights in neural network nodes. DTs are primarily employed for grouping tasks and are highly regarded as effective classification models in data mining. Each tree is composed of nodes and branches, with nodes representing features to be classified and branches defining possible values. Due to their analytical simplicity and accuracy across various data types, DTs are widely utilized in numerous fields. An example of a DT is shown in Fig. 6 [61].

3.3.3. Random forest

Ensemble learning involves creating multiple classifiers and combining their outputs. Popular techniques include Boosting and Bagging. In Boosting, extra weight is given to data points that were misclassified by previous classifiers. In contrast, Bagging constructs each tree independently using bootstrap samples and aggregates the predictions through majority voting. RF, depicted in Fig. 7, is an ensemble method consisting of numerous randomly created DTs. This algorithm introduces randomness by subsampling the training data and selecting node tests. Its robustness against overfitting and capability to manage complex features make it widely applicable in fields such as genetics, spectroscopy, and medical diagnostics.

3.3.4. LSBoost

Boosting is a popular technique in ensemble learning used for regression problems. It involves combining multiple weak models, such as DTs, to create a stronger predictive model. This method operates iteratively, with each new model focusing on the errors made by previous models to improve overall performance. The LSBoost regression algorithm utilizes the Least Squares Regression (LSR) model as its base and enhances predictions by weighting the misclassified data points. Boosting methods, including LSBoost, are highly effective and useful in data science and regression modeling due to their adaptability and robustness against noisy data (Fig. 8).

3.3.5. Modeling process of machine learning methods

Gene Expression Programming (GEP), DTs, RF, and LSBoost are prominent techniques in ML and data analysis, each offering unique capabilities and applications. GEP, grounded in GA, is used for symbolic regression and function discovery. It generates computer programs to tackle complex problems by integrating elements of GAs and GP. Initially, GEP represents solutions as linear chromosomes, which are then converted into tree-like structures.

DTs are used for both classification and regression tasks by partitioning the feature space into smaller regions using straightforward decision rules, making them easy to interpret and visualize. However, they can be prone to instability when dealing with noisy or complex datasets. In contrast, RFs are constructed by aggregating multiple DTs. By combining predictions from various trees through majority voting, RFs offer enhanced robustness against noisy data and are well-suited for managing high-dimensional spaces.

LSBoost is a boosting algorithm tailored for regression tasks, combining the benefits of LSR with boosting strategies. This algorithm iteratively adjusts the weights of data points, placing more emphasis on those that have been misclassified to enhance model performance. Generally, GEP is ideal for handling complex and symbolic problems, DTs are suited for more straightforward issues, RFs excel in scalability and robustness against noisy data, and LSBoost is effective for precise regression even in the presence of noise. The choice of method depends on the specific nature of the problem. Table 3 outlines the limitations of each machine learning method.

4. Discussion and results

The dataset was divided into two main sections: IFT and CA. Each section was further split into two subsets: 80 % for training and

Table 2
Statistical dataset of real and experimental extracted data.

|	Salinity (wt.%)	H_2 (%)	N_2 (%)	CH_4 (%)	CO_2 (%)	Pressure (psi)	Temperature (C°)	RO B (kg/m³)	RO MIX (kg/m³)	Delta RO (kg/m³)	CA ($\chi°$)	IFT (mN/m)
Max	20.000	80.000	70.000	70.000	40.000	3000.000	70.000	1099.600	181.740	1087.310	50.360	85.920
Min	2.000	20.000	5.000	5.000	5.000	500.000	30.000	986.540	8.140	833.600	22.200	26.640
Range	18.000	60.000	65.000	65.000	35.000	2500.000	40.000	113.060	173.600	253.710	28.160	59.280
Median	10.000	60.000	5.000	5.000	5.000	1750.000	50.000	1052.845	49.870	994.995	35.900	60.780
Mod	2.000	80.000	5.000	5.000	5.000	500.000	30.000	1009.620	11.100	995.950	32.100	62.840
Mean	10.400	54.000	14.333	21.333	10.333	1750.000	50.000	1049.693	57.436	992.256	35.993	59.183
variance	42.659	397.510	373.055	491.774	111.605	729490.885	200.089	1131.686	1319.751	2321.004	21.314	127.498
Skewness	0.167	−0.307	1.971	1.025	1.866	0.000	0.000	−0.178	0.959	−0.422	0.135	−0.470
kurtosis	−1.376	−1.062	2.457	−0.451	2.053	−1.269	−1.300	−1.312	0.539	−0.174	−0.324	−0.365
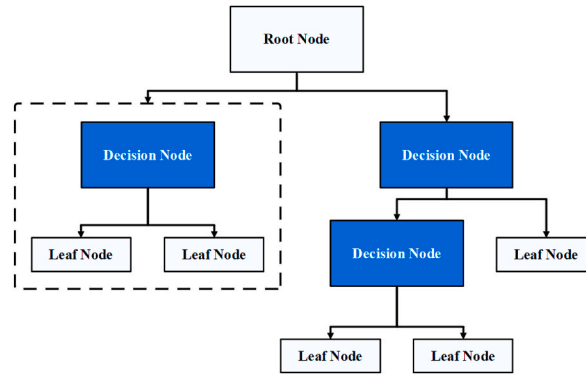
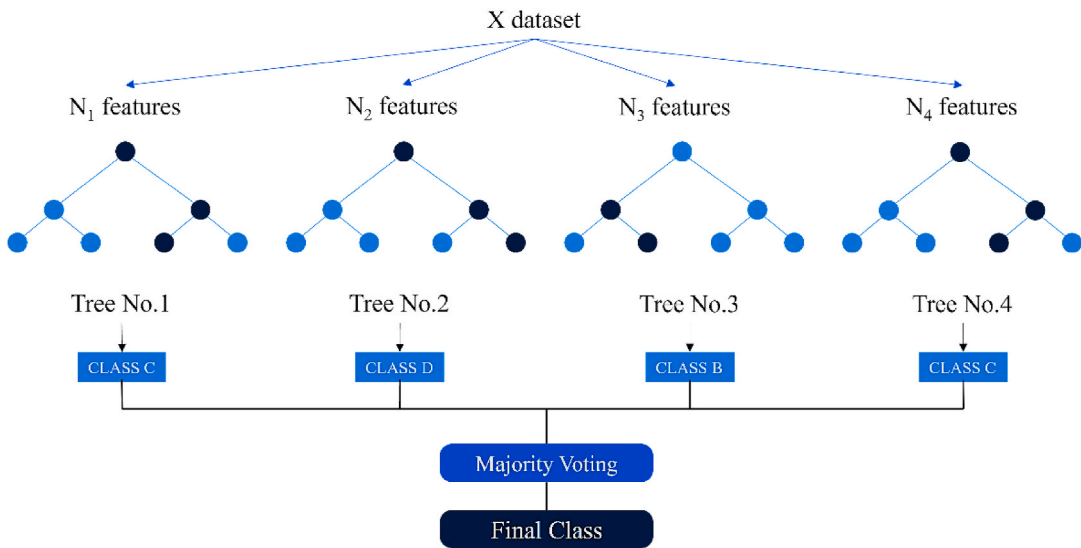**Fig. 6.** Schematic diagram of DT algorithm.

Fig. 7. Schematic diagram of RF algorithm.

20 % for testing. The training subset was used for model development and optimization, allowing the model to learn and establish relationships between parameters for accurate predictions. After the training process, the model was evaluated using the testing subset to assess its performance without further parameter adjustments. The model demonstrated satisfactory performance across both the training and testing datasets, with this partitioning approach aiding in the validation of the model's effectiveness.

This study emphasizes the significance of outlier removal for enhancing learning efficacy and achieving accurate model performance. The approach used to evaluate variable suitability and optimize the model incorporated standard deviation (SD). SD is a key measure for understanding data variability and distribution, and it assists in feature selection decisions. Additionally, it plays a crucial role in feature scaling, evaluating model stability, and identifying anomalous data points.

To identify and eliminate outliers, the three-sigma rule is applied, based on normal distribution principles: 68 % of data lies within ±1 SD, 95 % within ±2 SDs, and 99.7 % within ±3 SDs. Data points beyond three SDs are outliers and are removed. The process involves three steps: calculating the mean and SD, establishing the cutoff range, and identifying and removing outliers.

Initially, the mean and SD of the dataset were calculated.

$$\textbf{\textit{Mean}} \ (\pmb{\mu}) = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad\qquad 9$$

$$\textbf{\textit{Standard Deviation}} \ (\pmb{\sigma}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \qquad\qquad 10$$

In equations (9) and (10), N represents the number of data points, and x_i denotes each data point. Subsequently, the cutoff range for
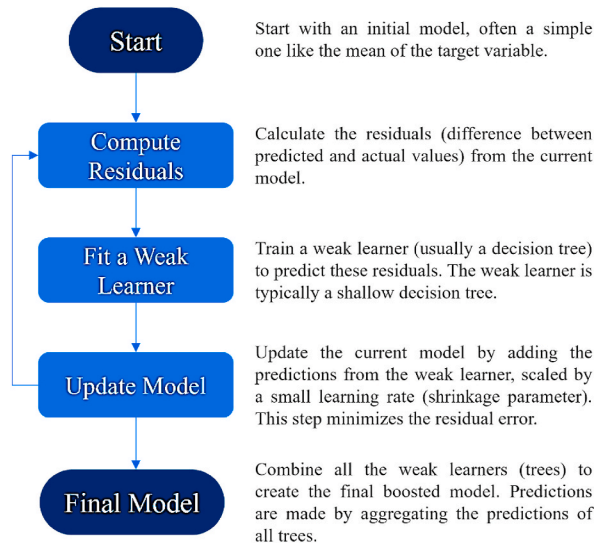
Fig. 8. Schematic diagram of LSBoost algorithm.

Table 3
GEP constant on correlations.

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
−164.7823	12805.1802	9.7586	7.2342	−13.2393	−2.2096	−2.5387	6.6887	10.5211	11.7410

outliers is identified, which is typically defined as the mean $\pm$ 3 SDs.

$$\textbf{\textit{Lower Bound}} = \boldsymbol{\mu} - \textbf{3}\boldsymbol{\sigma}$$

$$\textbf{\textit{Upper Bound}} = \boldsymbol{\mu} - \textbf{3}\boldsymbol{\sigma}$$

Any data point that falls outside this range is classified as an outlier and can be excluded from the dataset. This method is simple, easy to implement, and effective, especially for datasets that follow a normal distribution.

4.1. Hyperparameter optimization

To optimize hyperparameter of machine learning moldes, Bayesian and random search were employed. As shown in Fig. 9, after assigning initial values for hyperparameters, these values were updated for 30 iterations and at each step MSE of each model calculated. After that the model among them with the lowest MSE values was chosen as optimized model. Given the selected methods, the necessary explanations for interpreting each of them is provided in Table 4.

4.2. Model evaluation and performance

4.2.1. Evaluation metrics

Evaluation metrics are crucial in ML as they help measure, assess, and improve model performance. Choosing the right metrics is essential for ensuring the quality and effectiveness of ML models, as they provide critical insights into the models' accuracy and predictive ability on test data. Thus, the accurate selection of evaluation metrics is vital for the success of any ML project.

One key evaluation metric is the coefficient of R^2, which measures the proportion of variation in the dependent variable that is explained by the model. R^2 values range from 0 to 1, with higher values indicating a better fit. Additionally, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are important metrics for assessing model performance. RMSE quantifies the average squared differences between actual and predicted values, while MAE measures the average absolute differences. Lower values of RMSE and MAE signify higher model accuracy.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(IFT, CA_i^{exp} - IFT, CA_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(IFT, CA_i^{exp} - \overline{IFT, CA^{exp}}\right)^2}$$
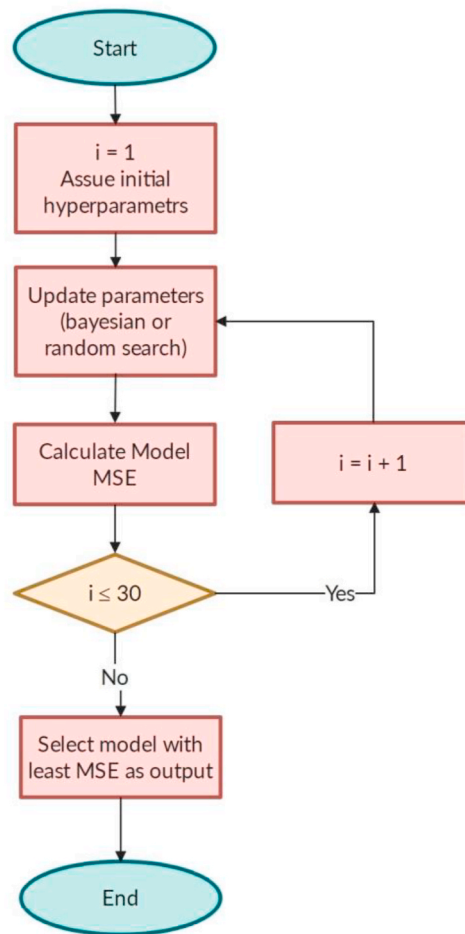
11

Fig. 9. Flowchart of hyperparameters tuni.

Table 4
Limitations of each machine learning method.

Machine Learning Method	Limitations
Genetic Expression Programming	**Need for Precise Tuning:** Requires careful tuning and complex parameters to achieve optimal models. **High Processing Time:** Due to simultaneous evolution of multiple models, it requires more processing time compared to other methods. **Unstable Modeling:** The final models may be unstable and sensitive to small changes in data.
Decision Tree	**Overfitting:** May become too dependent on the training data and fail to generalize well to new data if the data is complex. **Lack of Flexibility:** The tree structure may not capture complex relationships between features effectively. **Sensitivity to Small Changes:** Minor changes in data can lead to significant changes in the tree structure.
Random Forest	**Difficult Interpretation:** The final model can be complex due to the use of multiple decision trees, making it hard to interpret. **Need for Multiple Parameter Tuning:** Involves tuning parameters such as the number of trees and their depth, which can be complex. **High Processing Time:** The processing time increases with the number of trees, especially for large datasets.
Least Squares Boosting	**Sensitivity to Noise:** The model may become sensitive to noise in the data due to overfitting. **Model Complexity:** This method can create complex models that are difficult to interpret. **Need for Large Datasets:** Requires large and diverse datasets to model non-linear relationships effectively.

$$RMSE = \sqrt{\frac{\sum\limits_{i=1}^{N}\left(IFT,CA_i^{exp} - IFT,CA_i^{pred}\right)^2}{N}} \qquad\qquad 12$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| \left(IFT, CA_i^{exp} - IFT, CA_i^{pred} \right) \right| \qquad 13$$

Equations (11)–(13) involve the predicted values by the model (IFT, CA_i^{pred}), experimental values (IFT, CA_i^{exp}), and mean values ($\overline{IFT, CA^{exp}}$), as well as N, which represents the number of data points. Table 5 presents the calculated values and the reported errors in the data. Subsequently, the predictive capability of the model is evaluated through graphical results. The graphical assessment phase utilizes four distinct visualization techniques to demonstrate the accuracy of the developed models: cross plots, CFD plots, KDE plots, and the Hat Matrix.

4.2.2. Predictive power assessment

Table 6 shows the statistical values for training, testing, and overall datasets used to assess model accuracy using RMSE and R^2 metrics. RMSE provides an unbiased evaluation of performance, while R^2 indicates the model's ability to explain data variability. The LSBoost-BO model outperforms others in both metrics across all datasets for IFT and CA, with high R^2 and low RMSE values. It achieves the lowest RMSE of $0.413111°$ and the highest R^2 of 0.986999 for the CA dataset, demonstrating superior accuracy in predicting target variables (see Table 7).

The LSBoost algorithm is an ensemble method designed for regression problems, utilizing gradient boosting techniques to iteratively apply the base model and minimize prediction errors. By optimizing the least square's objective function, LSBoost aims to reduce the discrepancy between predicted and actual values. When optimized with Bayesian and RS methods, LSBoost achieves exceptional results. For the IFT dataset, the LSBoost model yields an RMSE of 0.717881 mN/m and an R^2 value of 0.995956, demonstrating its high accuracy and effectiveness.

Other models, such as RF and RT, also demonstrated good performance. The RF model, optimized using the Bayesian algorithm, achieved an RMSE of $0.609646°$ and an R^2 of 0.971686 for the CA dataset. For the IFT dataset, the RF-BO model calculated an RMSE of 1.147666 mN/m and an R^2 of 0.989665. Additionally, the RT and RF models optimized with Bayesian and RS algorithms exhibit a high density of predicted data points around the line X = Y, indicating high accuracy of the results (Figs. 10 and 11).

The kernel density estimation (KDE) plots shown in **Fig. 12** illustrate the distribution of residual errors for CA predictions across both the test and training datasets. The curves for LSBoost-BO, with peaks near zero, signify high prediction accuracy and lower variance in the residual errors. In contrast, the RT-BO curves, characterized by broad and shallow peaks, indicate higher variance and less accurate predictions. Additionally, the presence of longer tails in these curves suggests outliers or more significant prediction errors. It is also important to note that the residual error is calculated using the following equation (Equation (8)):

$$Re_i = y_i - \widehat{y}_i \qquad 14$$

Table 5
Provides detailed information on the hyperparameters used in developing the models (RT-RF-LSBoost) with both optimization techniques (Bayesian and RS).

Section	explanation	Reference
Bayesian algorithm	This research addresses the tuning of parameters in ML algorithms and the use of Bayesian optimization as an effective solution. Bayesian optimization, which outperforms other global optimization algorithms, operates on the assumption that the unknown function is sampled from a Gaussian process and uses criteria such as Expected Improvement (EI) or Upper Confidence Bound (UCB). This approach facilitates the automatic optimization of parameters. The study's results indicate that Bayesian optimization is not only effective in hyperparameter optimization but also highlights the significance of fully Bayesian treatment and the introduction of new algorithms to accommodate costs and concurrent experiments.	[62–64]
Random Search	RS Optimization is a versatile technique in ML used to optimally explore the hyperparameter space. Unlike systematic methods such as grid search that evaluate predefined combinations, RS selects hyperparameters randomly from specified ranges. This method can identify high-performance combinations that may be overlooked by other methods. RS is useful for extensive search spaces and limited computational resources and is accessible for trainees and professionals due to its simplicity and ease of implementation. This technique is also robust against noisy and discontinuous objective functions, allowing computational resources to be focused on valuable areas. The study shows that IFT and CA data have been meticulously examined using these methods, and the results are detailed comprehensively.	[23,65]
Hyperparameter Values and Evaluation	In this study, the selected hyperparameters for the models in RS and Bayesian optimization play a crucial role in optimizing the models for evaluating IFT and CA data. Detailed information about the hyperparameters and their evaluation in constructing the models (RT-RF-LSBoost) using these two optimization methods is provided. For the regression tree model, both optimization techniques have set the minimum leaf size to 1 to allow the tree to grow to its maximum depth. The RF model in RS is configured with 341 learners and a minimum leaf size of 1, while Bayesian optimization specifies 499 learners with the same minimum leaf size. In the IFT method, these numbers are 120 and 498, respectively. For the LSBoost model, RS in the CA method reports 29 learners, a minimum leaf size of 99, and a learning rate of 0.7290, whereas in the IFT method, it reports 158 learners, a minimum leaf size of 7, and a learning rate of 0.1323. In contrast, Bayesian optimization LSBoost in the CA method is configured with 98 learners, a minimum leaf size of 1, and a learning rate of 0.08, and in the IFT method, it is configured with 498 learners, a minimum leaf size of 1, and a learning rate of 0.4223. These differences in hyperparameters reflect different strategies for model optimization; RS allows models to grow to greater depths, while Bayesian optimization offers a more conservative and precise approach. These differences highlight the importance of hyperparameter tuning in ML to improve model accuracy, generalizability, and efficiency.	

Table 6

Hyperparameter information's chart.

Method	Model	Optimization method	
		Random search	Bayesian
Contact angle	Regression tree	Minimum leaf size = 1	Minimum leaf size = 1
	Random forest	Number of learners = 341	Number of learners = 499
		Minimum leaf size = 1	Minimum leaf size = 1
	LSBoost	Number of learners = 29	Number of learners = 98
		Minimum leaf size = 99	Minimum leaf size = 1
		Learn rate = 0.7290	Learn rate = 0.08
IFT	Regression tree	Minimum leaf size = 1	Minimum leaf size = 1
	Random forest	Number of learners = 120	Number of learners = 498
		Minimum leaf size = 2	Minimum leaf size = 1
	LSBoost	Number of learners = 158	Number of learners = 498
		Minimum leaf size = 7	Minimum leaf size = 1
		Learn rate = 0.1323	Learn rate = 0.4223

Table 7

Evaluation criteria of CA and IFT.

Method	Model	Statistical parameters	Train	Test	Total
CA	RT-BO	R^2	0.956563	0.890737	0.943269
		RMSE	0.754041	1.203805	0.862955
		MAE	0.58899	0.979949	0.667182
	RT-RS	R^2	0.956563	0.890737	0.943269
		RMSE	0.754041	1.203805	0.862955
		MAE	0.58899	0.979949	0.667182
	RF-BO	R^2	0.977133	0.95016	0.971686
		RMSE	0.547109	0.813037	0.609646
		MAE	0.410947	0.606992	0.450156
	RF-RS	R^2	0.976854	0.949266	0.971283
		RMSE	0.550435	0.820291	0.61397
		MAE	0.415237	0.613565	0.454903
	LSBoost-BO	R^2	0.988776	0.979971	0.986999
		RMSE	0.383295	0.515403	0.413111
		MAE	0.285649	0.386935	0.305906
	LSBoost-RS	R^2	0.965814	0.95061	0.962747
		RMSE	0.668947	0.809357	0.699288
		MAE	0.515703	0.626032	0.537769
IFT	RT-BO	R^2	0.981121	0.947058	0.975045
		RMSE	1.571916	2.45308	1.783327
		MAE	1.223674	1.927124	1.364364
	RT-RS	R^2	0.981121	0.947058	0.975045
		RMSE	1.571916	2.45308	1.783327
		MAE	1.223674	1.927124	1.364364
	RF-BO	R^2	0.992218	0.977904	0.989665
		RMSE	1.009224	1.584787	1.147666
		MAE	0.754388	1.160691	0.835649
	RF-RS	R^2	0.991917	0.977032	0.989262
		RMSE	1.028565	1.615753	1.169824
		MAE	0.769595	1.182686	0.852213
	LSBoost-BO	R^2	0.999074	0.996491	0.998614
		RMSE	0.348042	0.631536	0.420326
		MAE	0.266308	0.449785	0.303003
	LSBoost-RS	R^2	0.996478	0.993553	0.995956
		RMSE	0.678958	0.856057	0.717881
		MAE	0.523591	0.624443	0.543762

in this context, y_i represents the actual data points, and $\widehat{y}_i$ represents the predicted values for these data points. The curves similarly illustrate the distribution of residual errors as seen in the test set. Peaks near zero indicate good predictive performance on the training data. Comparing the curves of the training set with those of the test set, significantly taller and narrower curves in the training set may suggest overfitting, where the model performs well on the training data but less effectively on unseen test data. Conversely, if the curves are similar between the training and test sets, this indicates consistent model performance. The height, width, and tails of these curves offer insights into the model's accuracy and robustness, highlighting its ability to generalize from training data to test data.

Fig. 13 illustrates that the LSBoost-BO model successfully predicts over 97 % of CA values with an absolute relative error below 3 %. In comparison, the RF-RS and RF-BO models achieve this level of accuracy for approximately 90 % of CA values, while the RT-BO and RT-RS models achieve it for about 83 %. Fig. 14 demonstrates that the LSBoost-BO model also excels in predicting IFT values, with
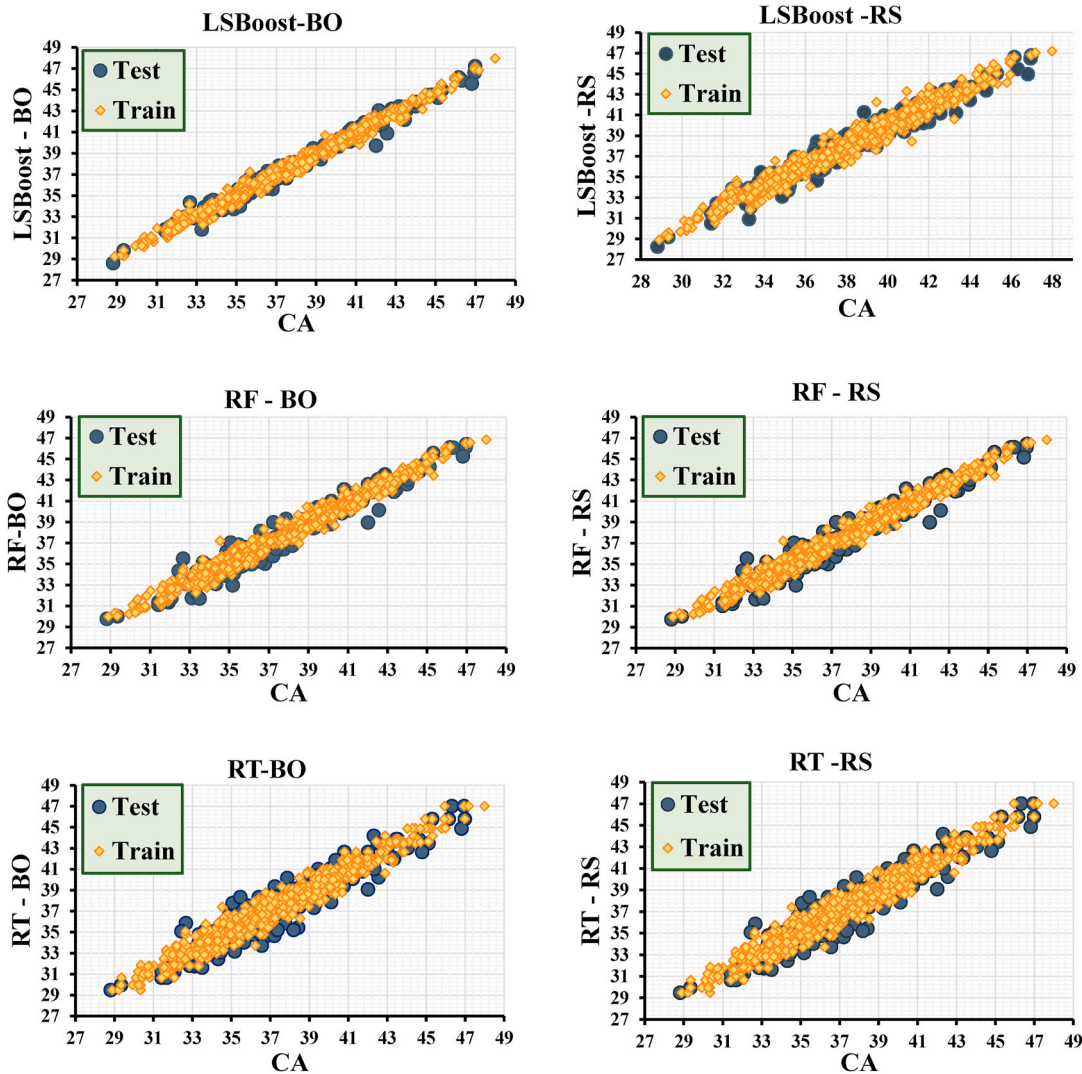
Fig. 10. Cross plots of predicted versus CA experimental data for different constructed models.

nearly 99 % of predictions having an absolute relative error of less than 2 %. The LSBoost-RS model predicts around 90 % of IFT values with an error under 2 % and CA values with an error under 3 %. The RF-RS and RF-BO models predict about 75 % of IFT values within a 2 % error margin and 90 % within a 3 % margin. In contrast, the RT-BO and RT-RS models show the least performance, with roughly 50 % of IFT predictions falling within a 2 % error margin and 70 % within a 3 % margin. Overall, all models demonstrate satisfactory performance with relatively low absolute relative error percentages.

Overall, the effectiveness of a ML algorithm is influenced by its underlying theoretical principles, the chosen architectural setup (e. g., tree-based structures), and the characteristics of the problem being addressed. Combining optimization techniques with ML approaches can enhance hyperparameter tuning and improve the model's overall performance. Additionally, using statistical metrics and graphical representations is an effective way to showcase the relative performance of different models.

Based on the above discussions, the proposed models can be ranked in the following order of accuracy:

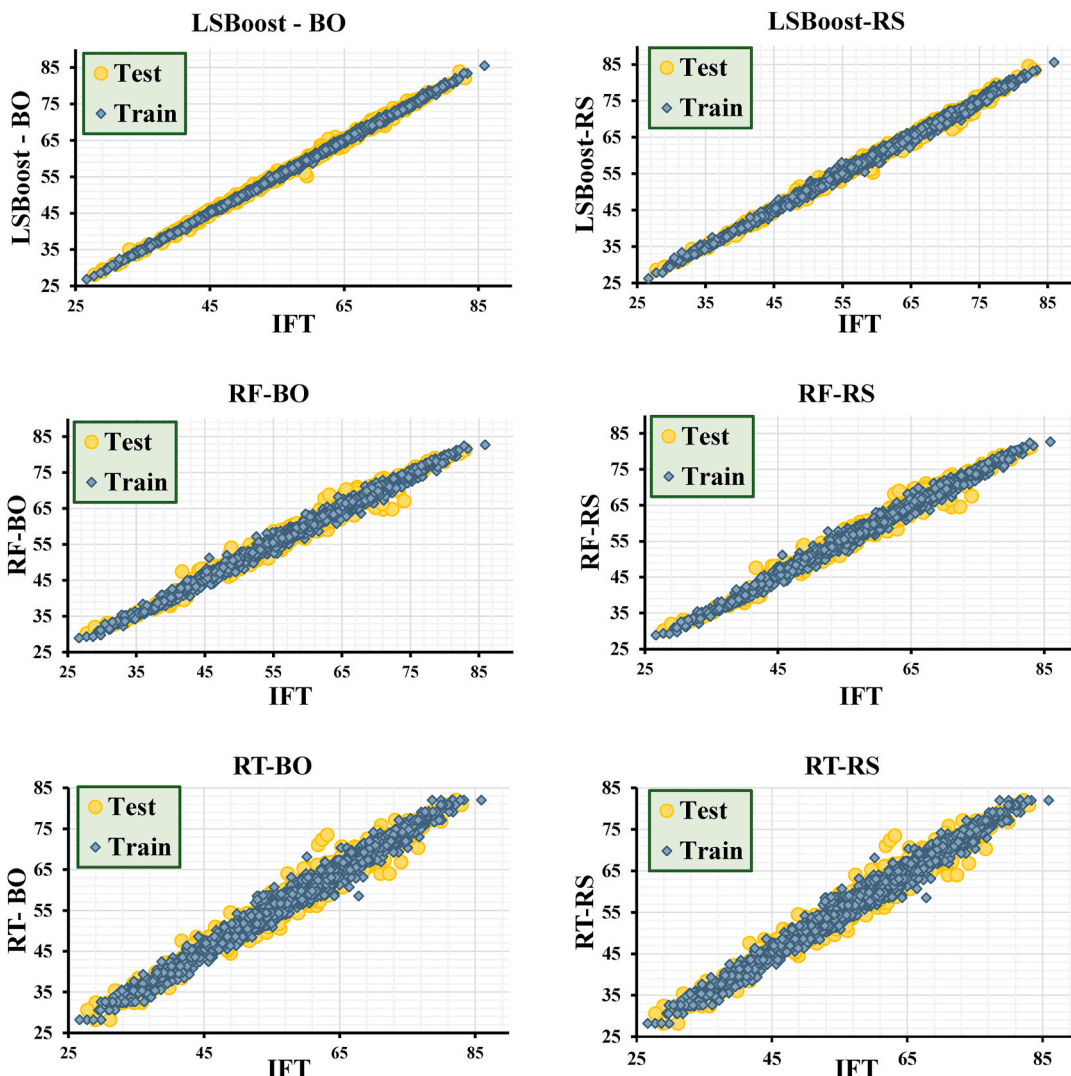| Number	Methods
1	LSBoost - BO
2	LSBoost - RS
3	RF - BO
4	RF - RS
5	RT - BO/RT - RS

Fig. 11. Cross plots of predicted versus IFT experimental data for different constructed models.

4.3. Removal of outlier data

Removing outliers or noisy data during preprocessing can significantly enhance the quality and performance of ML models in petroleum engineering. By filtering out these data points, the noise level is reduced, making the model less sensitive to anomalies and improving its accuracy and ability to generalize. This process also speeds up model training and conserves computational resources, as handling outliers often demands additional time and resources. Therefore, eliminating outliers not only improves model performance and reduces training time but also enhances computational efficiency and system optimization.

To identify outliers in a dataset, the Williams plot is used, which helps assess the reliability of the model (Fig. 15). This plot works by utilizing the Hat matrix and standardized residuals (SR). The Hat matrix assists in identifying leverage points, which are determined by the diagonal elements of the Hat matrix (H). If the leverage value of a point exceeds the leverage warning value (Hat*), calculated by the formula $\boldsymbol{Hat}^* = \frac{3(p+1)}{n}$, that point is considered a high leverage point. Standardized residuals represent the difference between the actual and predicted values of the model, expressed in a standardized form.

Points that fall outside the statistical safe ranges for the model ($0 \leq H \leq H^*$ and $-3 \leq SR \leq 3$) are categorized into three groups: vertical outliers, good leverage points, and bad leverage points. Vertical outliers have standardized residuals outside the statistical range, good leverage points help improve the model's accuracy, while bad leverage points harm the model and cause deviation in the results.

There are various methods for dealing with outliers, including removing outliers, adjusting outlier values to more reasonable ones, or using robust models that are less affected by outliers. In cases where there are missing data, imputation techniques can be applied, which might involve replacing missing values with the mean, median, or predicted values by the model. In some cases, additional data
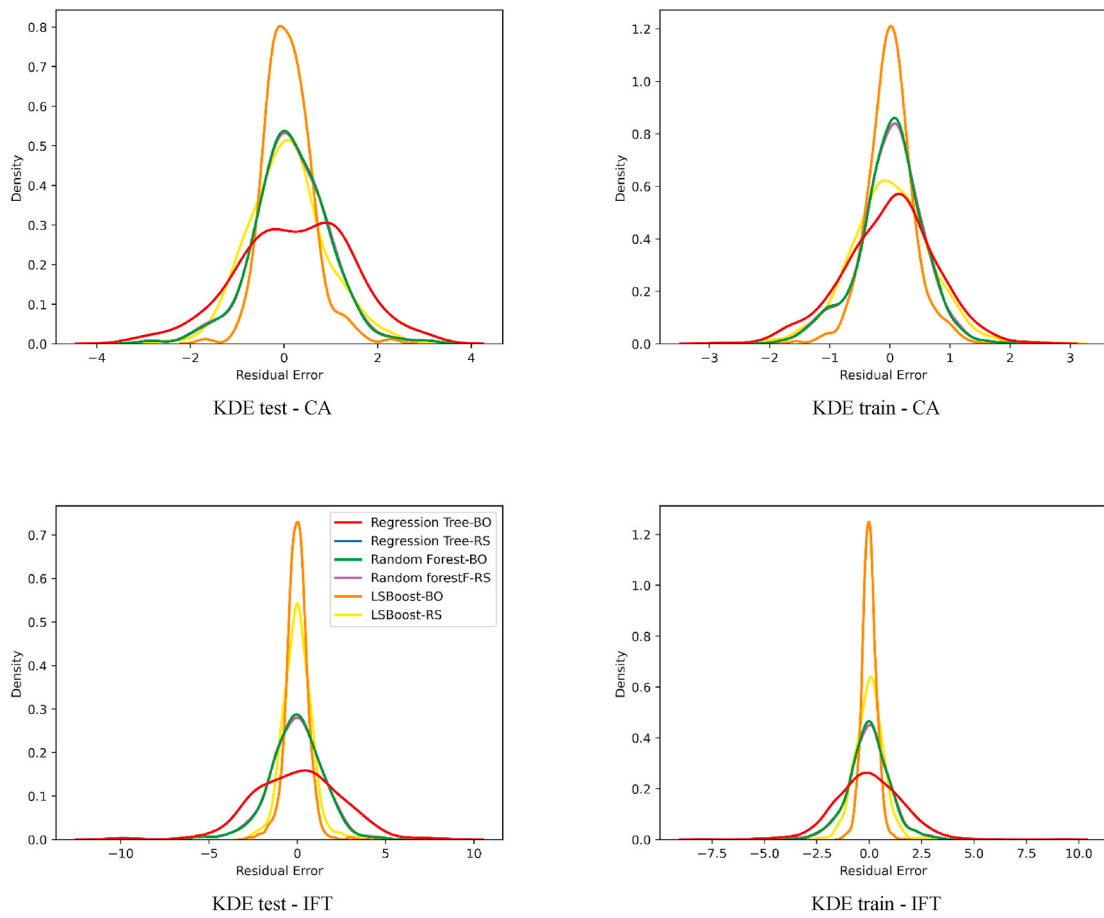
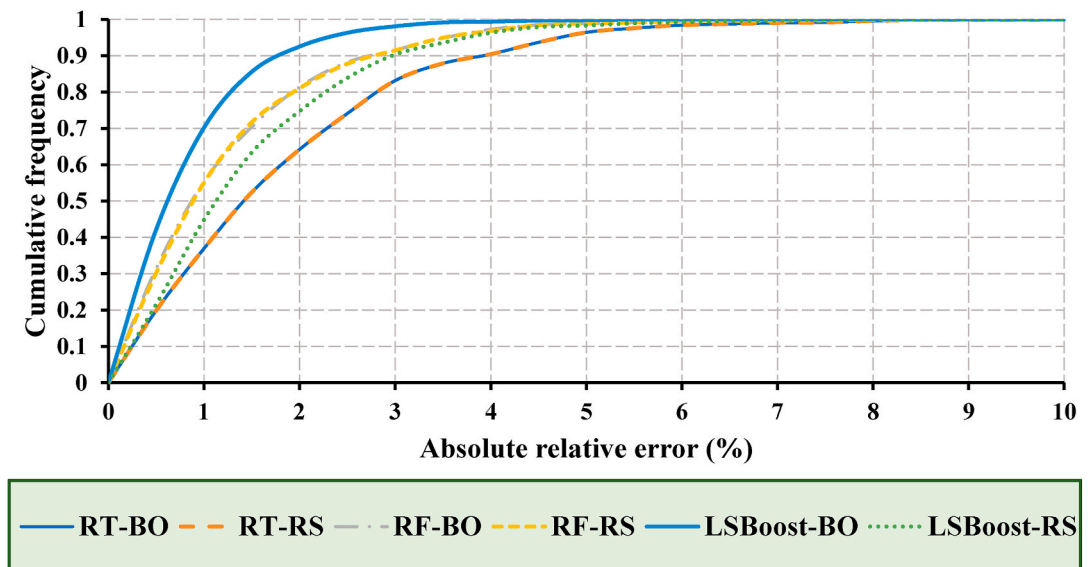**Fig. 12.** KDE plot of residual error of test and train datasets.
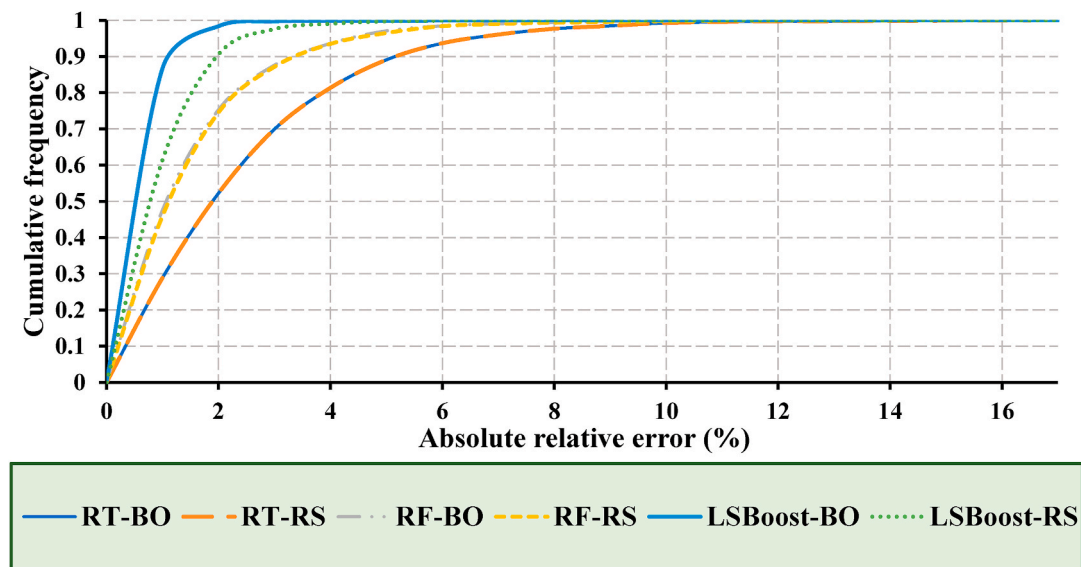

Fig. 13. C plot -CA.

Fig. 14. C plot -IFT.

may be required to fill in the missing values to ensure the model functions correctly.

5. Conclusions

In this study, advanced intelligent techniques, particularly LSBoost-BO, LSBoost-RS, RF-BO, RF-RS, RT-RS, and RT-BO, have been employed to model IFT in brine- H_2 systems with an inert gas. The various models utilized demonstrate distinct performances due to their diverse theories, architectures, and methodologies. Overall, evaluating and comparing these models can aid in selecting the most suitable modeling approach for the given problem. This study encompasses a well-organized and validated experimental dataset comprising 2,250 gas-water IFT data points and 950 CA data points from review studies, with IFT and CA as output parameters and inputs including temperature, pressure, density, weight percentage, and brine salinity. Statistical and graphical error analyses indicate satisfactory performance of the employed modeling techniques. (For CA, R^2 and RMSE values range from 0.943269 to 0.986999 and 0.413111°–0.862955°, respectively, and for IFT, R^2 and RMSE values range from 0.975045 to 0.998614 and 0.420326 mN/m to 1.783327 mN/m, respectively). However, the LSBoost-BO modeling approach demonstrated superior performance compared to other methods, with R^2 and RMSE values of 0.986999 and 0.413111° for CA, and 0.998614 and 0.420326 mN/m for IFT.

High R^2 values and low RMSE indicate accurate predictions that aid in precise design and optimization during CA and IFT storage phases. Accurate models are essential for risk analysis and project success. Furthermore, trend analysis has confirmed the predictive potential of the LSBoost-BO model, especially in relation to the significant input parameters mentioned, which play a critical role in the aforementioned processes. Based on performance analysis, the models are ranked accordingly in Table 8.

In this study, the Leverage approach was used to identify outliers, revealing that only a small fraction of the total CA and IFT datasets, approximately 1.2632 % and 1.1111 %, respectively, were classified as outliers. These findings validate the statistical reliability of the datasets and, consequently, support the credibility of the LSBoost-BO model.

Additionally, the results indicate that parameters such as temperature, pressure, weight percentage, and, notably, the density difference between the brine and gas mixture have a significant impact on CA and IFT values. In this context, the reported results and predicted values based on the laboratory data available in the references cited in this study are as follows:

- With increasing temperature, the values of CA, IFT, and density difference decrease.
- With increasing pressure, CA values increase while IFT and density difference values decrease.
- With increasing salinity, the values of CA, IFT, and density difference all increase.

Changes in interfacial tension (IFT) with different conditions can significantly impact various practical aspects of hydrogen storage. Here's a summary of potential effects:

- **Storage Efficiency and Capacity**
- **Leakage and Integrity**
- **Operational Stability**
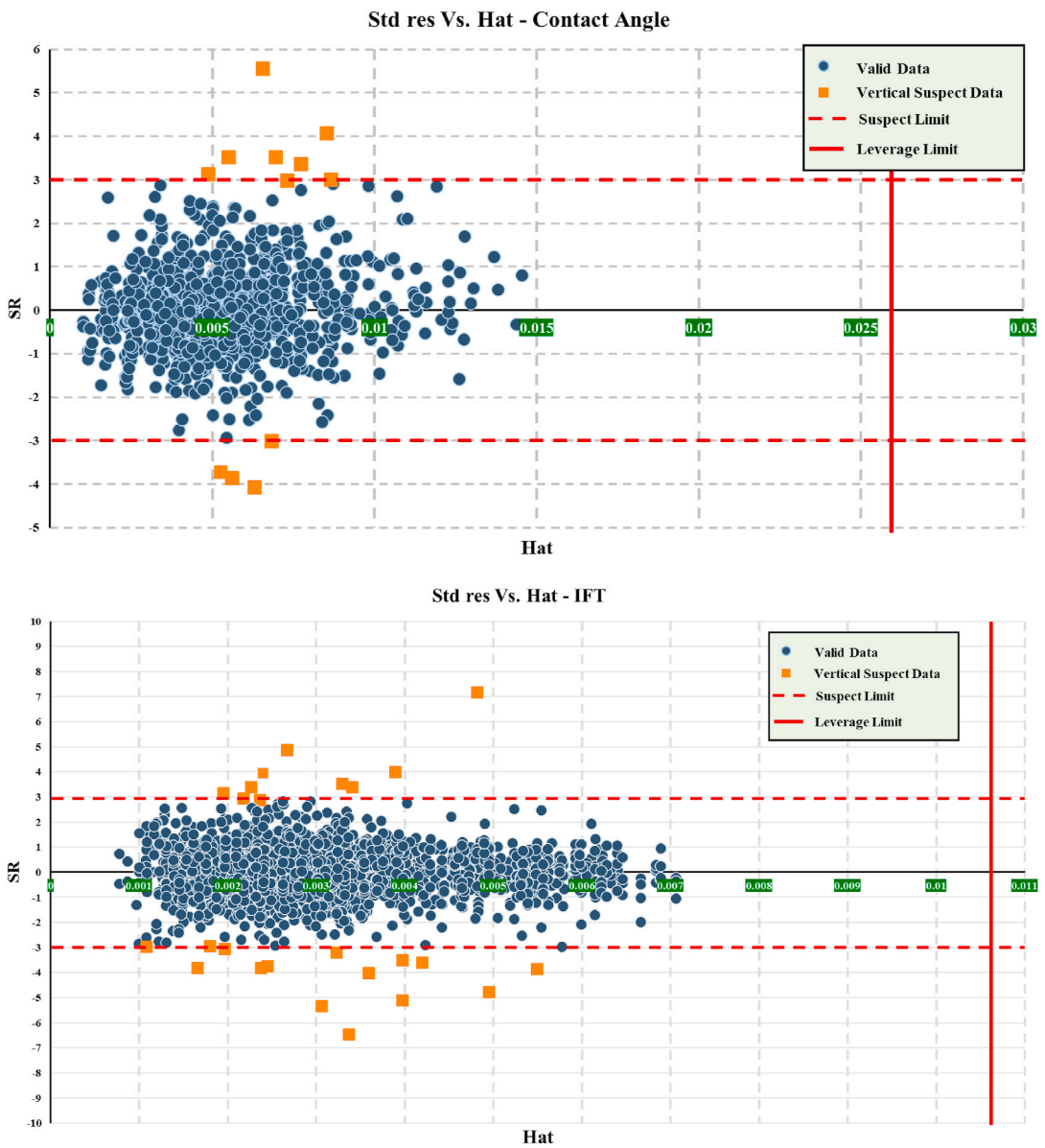- **Economic Considerations**

Fig. 15. Outlier detection using William's plot for LSBoost-BO model.

Table 8
Ranking and performance of ML methods for predicting CA and IFT in this study.

		R^2	RMSE
CA	LSBoost-BO	0.986999	0.413111
	RF-BO	0.971686	0.609646
	RF-RS	0.971283	0.61397
	LSBoost-RS	0.962747	0.699288
	RT-RS/RT-BO	0.943269	0.862955
IFT	LSBoost-BO	0.998614	0.420326
	RF-BO	0.995956	0.717881
	RF-RS	0.989665	1.147666
	LSBoost-RS	0.989262	1.169824
	RT-RS/RT-BO	0.975045	1.783327

Variations in interfacial tension (IFT) can significantly impact the stability and distribution of hydrogen within storage systems. Higher IFT may impede the effective dispersion of hydrogen, leading to reduced storage efficiency and capacity. Conversely, lower IFT can enhance the uniformity of hydrogen distribution, potentially improving storage density. Additionally, changes in IFT can affect the sealing and integrity of storage containers or geological formations. High IFT may lead to the formation of gas pockets that are prone to leakage, compromising the system's safety and reliability. Optimized IFT, however, can help maintain the integrity of the storage medium and reduce leakage risks.

Fluctuations in IFT also influence the stability of hydrogen storage operations, particularly with temperature and pressure changes. Suboptimal IFT conditions can result in operational challenges and safety concerns, while effective management of IFT can stabilize the storage environment and ensure consistent performance. The economic implications of IFT are significant, as inefficiencies or leakage due to improper IFT conditions can increase operational costs and reduce the economic viability of storage systems. Therefore, understanding and controlling IFT is crucial for optimizing hydrogen storage systems, enhancing their safety, reliability, and economic feasibility.

The sensitivity analysis reveals that temperature, specific gravity, and pressure are inversely related to IFT, meaning higher values of these parameters decrease IFT. Conversely, increased water salinity and density difference between brine and gas raise IFT. CA shows an inverse relationship with temperature and a direct relationship with pressure and water salinity. Higher temperature decreases both CA and IFT, while higher pressure increases CA and decreases IFT by changing the density difference. Higher water salinity increases CA, IFT, and density difference. The analysis aids in decision-making and optimization, with AI models like LSBoost-BO providing accurate IFT predictions and enhancing prediction accuracy in gas-water simulations.

CRediT authorship contribution statement

Mehdi Maleki: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mohammad Rasool Dehghani:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ali Akbari:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Yousef Kazemzadeh:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ali Ranjbar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

On behalf of the co-authors I declare no conflict of interest.

Abbreviations

AARD	Absolute Average Relative Deviation
CA	Contact Angle
CFD	Cumulative Frequency Distribution
CH_4	Methane
CO_2	Carbon Dioxide
DT	Decision Tree
EDA	Exploratory Data Analysis
EI	Expected Improvement
EOR	Enhanced Oil Recovery
GA	Genetic Algorithms
GBR	Gradient Boosting Regressor
GEP	Gene Expression Programming
GHG	Greenhouse Gas
GP	Genetic Programming
H_2	Hydrogen
IFT	Interfacial Tension
IQR	Interquartile Range
KDE	Kernel Density Estimate
LGB	Light Gradient Boosting
LMA	Levenberg-Marquardt Algorithm
LSBoost	Least Squares Boosting
LSR	Least Squares Regression
MAPE	Mean Absolute Percentage Error
ML	Machine Learning

(continued)

MLP	Multi-Layer Perceptron
P	Pressure
Pce	Capillary Entry Pressure
R^2	Determination
RF	Random Forest
RMSE	Root Mean Square Error
RS	Random Search
RT	Regression Tree
SD	Standard Deviation
ST	Surface Tension
SVM	Support Vector Machines
T	Temperature
UCB	Upper Confidence Bound
UGS	Underground Gas Storage
UHS	Underground Hydrogen Storage
XGBoost	Extreme Gradient Boosting

References

[1] N.S. Muhammed, B. Haq, D.J. Al Shehri, CO2 Rich Cushion Gas for Hydrogen Storage in Depleted Gas Reservoirs: Insight on Contact Angle and Surface Tension, vol. 50, 2024, pp. 1281–1301.

[2] S. Ding, Y. Xi, H. Jiang, G.J.A.G. Liu, CO2 Storage Capacity Estimation in Oil Reservoirs by Solubility and Mineral Trapping, vol. 89, 2018, pp. 121–128.

[3] M. Maleki, Y. Kazemzadeh, A. Dehghan Monfared, A. Hasan-Zadeh, S.J.T.C. Abbasi, Bio-enhanced oil recovery (BEOR) methods: All-important review of the occasions and challenges 102 (2024) 2364–2390.

[4] M. Alruqi, P. Sharma, B. Deepanraj, F.J.F. Shaik, Renewable energy approach towards powering the CI engine with ternary blends of algal biodiesel-diesel-diethyl ether: Bayesian optimized Gaussian process regression for modeling-optimization 334 (2023) 126827.

[5] M. Ali, N. Yekeen, N. Pal, A. Keshavarz, S. Iglauer, H.J.E.R. Hoteit, Influence of Pressure, Temperature and Organic Surface Concentration on Hydrogen Wettability of Caprock; Implications for Hydrogen Geo-Storage, vol. 7, 2021, pp. 5988–5996.

[6] I. Dincer, M. I. J. E. C. Aydin, and management, New paradigms in sustainable energy systems with hydrogen 283 (2023) 116950.

[7] M. Hosseini, J. Fahimpour, M. Ali, A. Keshavarz, S.J. J.o.C. Iglauer, I. Science, Hydrogen Wettability of Carbonate Formations: Implications for Hydrogen Geo-Storage, vol. 614, 2022, pp. 256–266.

[8] P. Sharma, et al., Application of modern approaches to the synthesis of biohydrogen from organic waste 48 (55) (2023) 21189–21213.

[9] M.R. Dehghani, S.F. Ghazi, Y.J.S.R. Kazemzadeh, Interfacial tension and wettability alteration during hydrogen and carbon dioxide storage in depleted gas reservoirs 14 (1) (2024) 11594.

[10] A.A. Abd, M.R. Othman, H.S. Majdi, Z.J.R.E. Helwani, Green Route for Biomethane and Hydrogen Production via Integration of Biogas Upgrading Using Pressure Swing Adsorption and Steam-Methane Reforming Process, vol. 210, 2023, pp. 64–78.

[11] L. Hashemi, W. Glerum, R. Farajzadeh, H.J. Hajibeygi, Contact Angle Measurement for Hydrogen/brine/sandstone System Using Captive-Bubble Method Relevant for Underground Hydrogen Storage, vol. 154, 2021 103964.

[12] P. Gabrielli, et al., Seasonal Energy Storage for Zero-Emissions Multi-Energy Systems via Underground Hydrogen Storage, vol. 121, 2020 109629.

[13] H.B. Stone, I. Veldhuis, R. N. J. G. S. Richardson, London, Special Publications, Underground hydrogen storage in the UK 313 (1) (2009) 217–226.

[14] T. Rudolph, Underground hydrogen storage–current developments and opportunities, in: EAGE/DGMK Joint Workshop on Underground Storage of Hydrogen, vol. 2019, European Association of Geoscientists & Engineers, 2019, pp. 1–2, no. 1.

[15] A. Al-Yaseri, et al., Assessment of CO2/shale Interfacial Tension, vol. 627, 2021 127118.

[16] N. Heinemann, et al., Hydrogen storage in saline aquifers: the role of cushion gas for injection and production 46 (79) (2021) 39284–39296.

[17] A. Keshavarz, H. Abid, M. Ali, S.J. J.o.C. Iglauer, I. Science, Hydrogen Diffusion in Coal: Implications for Hydrogen Geo-storage, vol. 608, 2022, pp. 1457–1462.

[18] M. Yue, et al., Hydrogen Energy Systems: A Critical Review of Technologies, Applications, Trends and Challenges, vol. 146, 2021 111180.

[19] Y. Balali, S.J.R. Stegen, S.E. Reviews, Review of energy storage systems for vehicles based on technology, environmental impacts, and costs 135 (2021) 110185.

[20] A. Lemieux, A. Shkarupin, K.J. Sharp, Geologic feasibility of underground hydrogen storage in Canada 45 (56) (2020) 32243–32259.

[21] J. Mouli-Castillo, N. Heinemann, K.J.A.E. Edlmann, Mapping Geological Hydrogen Storage Capacity and Regional Heating Demands: an Applied UK Case Study, vol. 283, 2021 116348.

[22] T. Van de Graaf, I. Overland, D. Scholten, K.J.E. Westphal, S. Science, The New Oil? the Geopolitics and International Governance of Hydrogen, vol. 70, 2020 101667.

[23] R. Kalule, H.A. Abderrahmane, W. Alameri, M.J.S.R. Sassi, Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs 13 (1) (2023) 9855.

[24] M.R. Dehghani, S.F. Ghazi, Y. Kazemzadeh, Interfacial tension and wettability alteration during hydrogen and carbon dioxide storage in depleted gas reservoirs, Sci. Rep. 14 (1) (2024) 11594.

[25] U. Eberle, G. Arnold, R.J. Von Helmolt, Hydrogen storage in metal–hydrogen systems and their derivatives 154 (2) (2006) 456–460.

[26] R. Taccani, S. Malabotti, C. Dall'Armi, D.J.I.S.P. Micheli, High energy density storage of gaseous marine fuels: an innovative concept and its application to a hydrogen powered ferry 67 (1) (2020) 33–56.

[27] R.J.R. Tarkowski, S.E. Reviews, Underground Hydrogen Storage: Characteristics and Prospects, vol. 105, 2019, pp. 86–94.

[28] M. Panfilov, Underground and pipeline hydrogen storage, in: Compendium of Hydrogen Energy, Elsevier, 2016, pp. 91–115.

[29] B. Nemati, et al., A Sustainable Approach for Site Selection of Underground Hydrogen Storage Facilities Using Fuzzy-Delphi Methodology 6 (2020) 5–16.

[30] A. Ebrahimiyekta, Characterization of Geochemical Interactions and Migration of Hydrogen in Sandstone Sedimentary Formations: Application to Geological Storage, Université d'Orléans, 2017.

[31] N. Dopffel, S. Jansen, J.J. Gerritse, Microbial side effects of underground hydrogen storage–Knowledge gaps, risks and opportunities for successful implementation 46 (12) (2021) 8594–8606.

[32] A. Aftab, A. Hassanpouryouzband, Q. Xie, L.L. Machuca, M.J.I. Sarmadivaleh, E.C. Research, Toward a fundamental understanding of geological hydrogen storage 61 (9) (2022) 3233–3253.

[33] E.I. Epelle, et al., Perspectives and prospects of underground hydrogen storage and natural hydrogen 6 (14) (2022) 3324–3343.

[34] C. Benetatos, et al., How underground systems can contribute to meet the challenges of energy transition 58 (2–3) (2021) 65–80.

[35] F. Feldmann, B. Hagemann, L. Ganzer, M.J.E.E.S. Panfilov, Numerical Simulation of Hydrodynamic and Gas Mixing Processes in Underground Hydrogen Storages, vol. 75, 2016, pp. 1–15.

[36] C. Aggelopoulos, M. Robin, O. Vizika, Interfacial tension between CO2 and brine (NaCl+ CaCl2) at elevated pressures and temperatures: the additive effect of different salts 34 (4) (2011) 505–511.

[37] T. Akamine, T. Tosuai, R. Ramadhan, N. Promsuk, F. Srisuriyachai, S. Tangparitkul, Visualizing oil displacement by nanofluids at pore scale: a concentration-dependent nanofluid spreading induced by structural disjoining pressure, Capillarity 12 (1) (2024) 17–26.

[38] S. Tangparitkul, A. Sukee, J. Jiang, D. Harbottle, Crude oil displacement enhanced by interfacially active nanoparticles and their coupling effect with low-salinity brines, J. Mol. Liq. 408 (2024) 125362.

[39] M. Hosseini, Y. Leonenko, Hydrogen-brine interfacial tension at subsurface conditions: implication for hydrogen geo-storage, in: 84th EAGE Annual Conference & Exhibition, vol. 2023, European Association of Geoscientists & Engineers, 2023, pp. 1–5, no. 1.

[40] C.S.W. Ng, H. Djema, M.N. Amar, A. J. J. i. j. o. h. e. Ghahfarokhi, Modeling interfacial tension of the hydrogen-brine system using robust machine learning techniques: implication for underground hydrogen storage 47 (93) (2022) 39595–39605.

[41] M. Behnamnia, N. Mozafari, A.D.J. Monfared, Rigorous Hybrid Machine Learning Approaches for Interfacial Tension Modeling in Brine-Hydrogen/cushion Gas Systems: Implication for Hydrogen Geo-Storage in the Presence of Cushion Gas, vol. 73, 2023 108995.

[42] N. Suetrong, T. Tosuai, H. Vo Thanh, W. Chantapakul, S. Tangparitkul, N. Promsuk, Predicting dynamic contact angle in immiscible fluid displacement: a machine learning approach for subsurface flow applications, Energy & Fuels 38 (5) (2024) 3635–3644.

[43] H.V. Thanh, H. Zhang, Z. Dai, T. Zhang, S. Tangparitkul, B. Min, Data-driven machine learning models for the prediction of hydrogen solubility in aqueous systems of varying salinity: implications for underground hydrogen storage, Int. J. Hydrogen Energy 55 (2024) 1422–1433.

[44] M. Masoudi, A. Hassanpouryouzband, H. Hellevang, R.S. Haszeldine, Lined rock caverns: a hydrogen storage solution, J. Energy Storage 84 (2024) 110927.

[45] A. Hassanpouryouzband, M. Wilkinson, R.S. Haszeldine, Hydrogen energy futures–foraging or farming? Chem. Soc. Rev. 53 (2024) 2258–2263.

[46] C.S.W. Ng, H. Djema, M.N. Amar, A.J. Ghahfarokhi, Modeling interfacial tension of the hydrogen-brine system using robust machine learning techniques: implication for underground hydrogen storage, Int. J. Hydrogen Energy 47 (93) (2022) 39595–39605.

[47] H.V. Thanh, M. Rahimi, Z. Dai, H. Zhang, T. Zhang, Predicting the wettability rocks/minerals-brine-hydrogen system for hydrogen storage: Re-evaluation approach by multi-machine learning scheme, Fuel 345 (2023) 128183.

[48] A. Ahmed, D.J. Siegel, Predicting hydrogen storage in MOFs via machine learning, Patterns 2 (7) (2021).

[49] S. Mao, et al., Efficient prediction of hydrogen storage performance in depleted gas reservoirs using machine learning, Appl. Energy 361 (2024) 122914.

[50] S. Mao, B. Chen, M. Morales, M. Malki, M. Mehana, Cushion gas effects on hydrogen storage in porous rocks: insights from reservoir simulation and deep learning, Int. J. Hydrogen Energy 68 (2024) 1033–1047.

[51] M. Kanaani, B. Sedaee, M. Asadian-Pakfar, M. Gilavand, Z. Almahmoudi, Development of multi-objective co-optimization framework for underground hydrogen storage and carbon dioxide storage using machine learning algorithms, J. Clean. Prod. 386 (2023) 135785.

[52] J.O. Helland, H.A. Friis, M. Assadi, Ł. Klimkowski, S. Nagy, Prediction of optimal production time during underground CH4 storage with cushion CO2 using reservoir simulations and artificial neural networks, Energy & Fuels 37 (23) (2023) 19022–19038.

[53] S. Bakhshian, A. Shariat, A. Raza, Assessment of CO2 storage potential in reservoirs with residual gas using deep learning, Interpretation 10 (3) (2022) SG11–SG20.

[54] H.V. Thanh, M. Safaei-Farouji, N. Wei, S.S. Band, A. Mosavi, Knowledge-based rigorous machine learning techniques to predict the deliverability of underground natural gas storage sites for contributing to sustainable development goals, Energy Rep. 8 (2022) 7643–7656.

[55] H. Wang, S. Williams-Stroud, D. Crandall, C. Chen, Machine learning and deep learning for mineralogy interpretation and CO2 saturation estimation in geological carbon Storage: a case study in the Illinois Basin, Fuel 361 (2024) 130586.

[56] M. Ali, Z. Tariq, M. Mubashir, M.S. Kamal, B. Yan, H. Hoteit, Prediction of pure mineral-H2-brine wettability using data-driven machine learning modeling: implications for H2 geo-storage, in: International Petroleum Technology Conference, IPTC, 2024. D031S097R007.

[57] M. Xie, M. Zhang, Z. Jin, Machine learning-based interfacial tension equations for (H2+ CO2)-water/brine systems over a wide range of temperature and pressure, Langmuir 40 (10) (2024) 5369–5377.

[58] N.S. Muhammed, B. Haq, D.J. Al Shehri, Role of methane as a cushion gas for hydrogen storage in depleted gas reservoirs 48 (76) (2023) 29663–29681.

[59] N.S. Muhammed, B. Haq, D.A.J. Al Shehri, Hydrogen storage in depleted gas reservoirs using nitrogen cushion gas: a contact angle and surface tension study 48 (98) (2023) 38782–38807.

[60] M.R. Dehghani, et al., Data driven models for predicting pH of CO2 in aqueous solutions: implications for CO2 sequestration, Results in Engineering (2024) 102889.

[61] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer, 2015.

[62] D.R.J. Jones, A Taxonomy of Global Optimization Methods Based on Response Surfaces, vol. 21, 2001, pp. 345–383.

[63] A.D.J. Bull, Convergence rates of efficient global optimization algorithms 12 (10) (2011).

[64] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian Optimization of Machine Learning Algorithms, vol. 25, 2012.

[65] A. Yasmin, W. Haider Butt, A.J. P.o. Daud, Ensemble effort estimation with metaheuristic hyperparameters and weight optimization for achieving accuracy 19 (4) (2024) e0300296.