

Machine Learning Models for Predicting Hydrogen Solubility in Saline Aqueous Solutions for Underground Storage

Sandrely Pereira da Silva^{1,2,5}
Caio Bezerra Souto Maior^{1,3,4,6}
Isis Didier Lins^{1,2,7}
Márcio José das Chagas Moura^{1,2,8}

¹Center for Risk Analysis, Reliability Engineering and Environmental Modeling (CEERMA)
Av. da Arquitetura, 211-351 - Cidade Universitária, Recife - PE, 50740-550

²Universidade Federal de Pernambuco (UFPE)
Av. da Arquitetura, s/n - Cidade Universitária, Recife - PE, 50740-550

³Technology Center – UFPE
Av. Marielle Franco, s/n - Km 59 - Nova, PE, 55014-900

⁴The B. John Garrick Institute for the Risk Sciences, University of California
Los Angeles, CA 90095, EUA

⁵sandrely.pereira@ufpe.br, ⁶caio.maior@ufpe.br, ⁷isis.lins@ufpe.br,
⁸marcio.cmoura@ufpe.br,

ABSTRACT

The large-scale adoption of underground hydrogen storage (UHS) is essential to transition to low-carbon energy systems. However, it remains constrained by technical challenges, including the accurate prediction of hydrogen solubility under varying subsurface conditions. This study explores the application of machine learning (ML) techniques to model hydrogen solubility in saline aqueous solutions, aiming to overcome the limitations of traditional thermodynamic models. A comprehensive methodology was developed, incorporating data preprocessing, feature selection, and the evaluation of multiple regression models, including CatBoost, XGBoost, LightGBM, and Support Vector Regressor. Among these, CatBoost demonstrated superior performance with an R^2 of 0.9939 and RMSE of 0.0188. These findings demonstrate the viability of ML models in providing accurate and scalable predictive tools, contributing to risk assessment in UHS projects. Ultimately, the proposed approach supports safe and efficient hydrogen storage in transitioning to sustainable energy systems.

KEYWORDS. Hydrogen Solubility. Underground Hydrogen Storage. Machine Learning.

Paper topics. EN&OG. EST&AM. AS&DS. IA.

1. Introduction

In the face of accelerating climate change, the urgent need to transition towards low-carbon energy systems has become increasingly evident. The global commitment to reducing greenhouse gas emissions has driven the diversification of the energy matrix, motivating investments in renewable and sustainable energy sources [Hassan et al. 2024]. Among these alternatives, hydrogen is important in decarbonizing various sectors, offering a clean energy carrier with high energy density and versatile applications across industries and transportation [Zhang et al. 2024]. Expanding the use of hydrogen is a strategic necessity to achieve long-term climate goals, and a technological challenge.

The large-scale utilization of hydrogen requires overcoming significant barriers related to its storage, distribution, and safety. Traditional storage methods, such as compressed gas and cryogenic liquid storage, face economic and technical limitations, particularly when considering the infrastructure needed for widespread adoption [Semedo et al. 2025]. As a promising alternative, underground hydrogen storage (UHS), including storage in depleted reservoirs, aquifers, or salt caverns, has gained attention due to its potential to accommodate large volumes and balance seasonal supply and demand [Ramesh Kumar et al. 2023]. Nevertheless, underground storage introduces new complexities, notably the interactions between hydrogen and the geological environment, including its solubility in brine and other formation fluids [Gbadamosi et al. 2023]. Predicting the solubility of hydrogen in aqueous solutions under various pressure, temperature, and salinity conditions is a promising approach to assess risks such as gas loss, microbiological activity, and geochemical reactions that may compromise storage efficiency and safety [Dehghani et al. 2024].

In this context, data-driven strategies, particularly machine learning (ML) techniques, are powerful tools to model complex physicochemical phenomena where traditional theoretical or empirical models may fall short [Keith et al. 2021]. Predicting hydrogen solubility in aqueous solutions involves non-linear interactions that are difficult to capture with classical thermodynamic equations, especially when accounting for the heterogeneity of underground environments [Mwikipunda et al. 2024]. Moreover, experimental measurements of hydrogen solubility at high pressures and diverse salinities are scarce, costly, and sometimes logistically unfeasible, leading to data gaps that hinder comprehensive modeling using conventional methods [Li et al. 2018]. Thus, ML offers advantages in addressing these challenges. ML models can learn directly from data without requiring explicit assumptions about functional forms or variable interactions, making them well-suited for complex or poorly understood systems [Maier et al. 2025]. They can capture non-linear factors influencing hydrogen solubility under varying subsurface conditions. They outperform traditional approaches by generalizing across diverse inputs. Once trained, they offer scalability and can be easily updated, making them well-suited for practical UHS applications [Zhou et al. 2022].

This study discusses the construction of a data-driven approach utilizing different ML regression models, such as CatBoost, XGBoost, Gradient Boosting, LightGBM, Linear Regression, Ridge Regression, and Support Vector Regressor for predicting hydrogen solubility in aqueous systems. The methodology applied aims to provide precise predictions and contribute to establishing UHS as a feasible and resilient option in the transition to renewable energy. The research followed a structured approach comprising several steps. Section 2 outlines the methodology utilized, including an Exploratory Data Analysis (EDA) to understand the underlying patterns and distributions in the dataset. Next, a preprocessing pipeline was applied, involving feature engineering, data splitting, applying the Yeo-Johnson transformation, standardizing the features, and performing variable selection using the stepwise method. Following preprocessing, the predictive models were implemented. In Section 3, the study presents and discusses the results of the modeling, highlighting the performance of each model and the implications for UHS. Lastly, Section 4 offers concluding observations.

2. Methodology

The methodology of the present study consists of exploratory data analysis, data preprocessing, modelling, and result analysis. The methodology flowchart can be visualized in Figure 1.

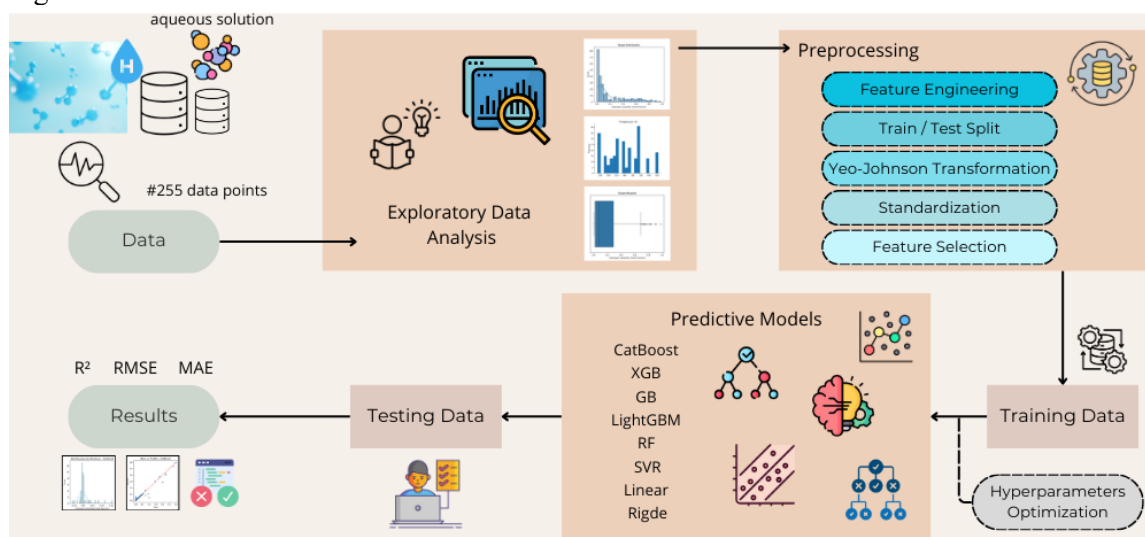


Figure 1: Methodology flowchart.

2.1. Exploratory data analysis

A first exploratory data analysis was conducted to understand the characteristics and distribution of the dataset variables. Thus, descriptive statistics were calculated for pressure, temperature, salinity, and hydrogen solubility, revealing that pressure and temperature exhibited high variability, with wide ranges across observations. Salinity values were mostly concentrated at lower percentages, and hydrogen solubility was skewed towards lower values, with a mean of approximately 0.190. These findings is presented in Table 1.

Table 1: Descriptive statistics.

Variable	count	mean	std	min	25%	50%	75%	max
Pressure (bar)	255	44.57	45.22	4.60	15.00	25.00	66.63	229.72
Temperature (K)	255	336.95	42.26	273.15	303.15	333.15	372.75	423.15
Salinity (% by weight)	255	1.44	1.80	0.00	0.00	1.00	3.00	5.00
Hydrogen solubility (mole fraction)	255	0.19	0.24	0.00	0.02	0.07	0.29	0.98

The dataset consists of 255 data points. Moreover, the missing values analysis indicated no NaN values across any columns, allowing the dataset to proceed without imputation. Additionally, non-informative variables, such as “Reference” and “Reference DOI”, were removed since they do not contribute to the modeling process.

Histograms showed that hydrogen solubility values were concentrated near zero, with a long tail toward higher solubility, confirming a right-skewed distribution. Similar right-skewness was observed for pressure and salinity distributions, while temperature followed a more uniform spread, as presented in Figure 2.

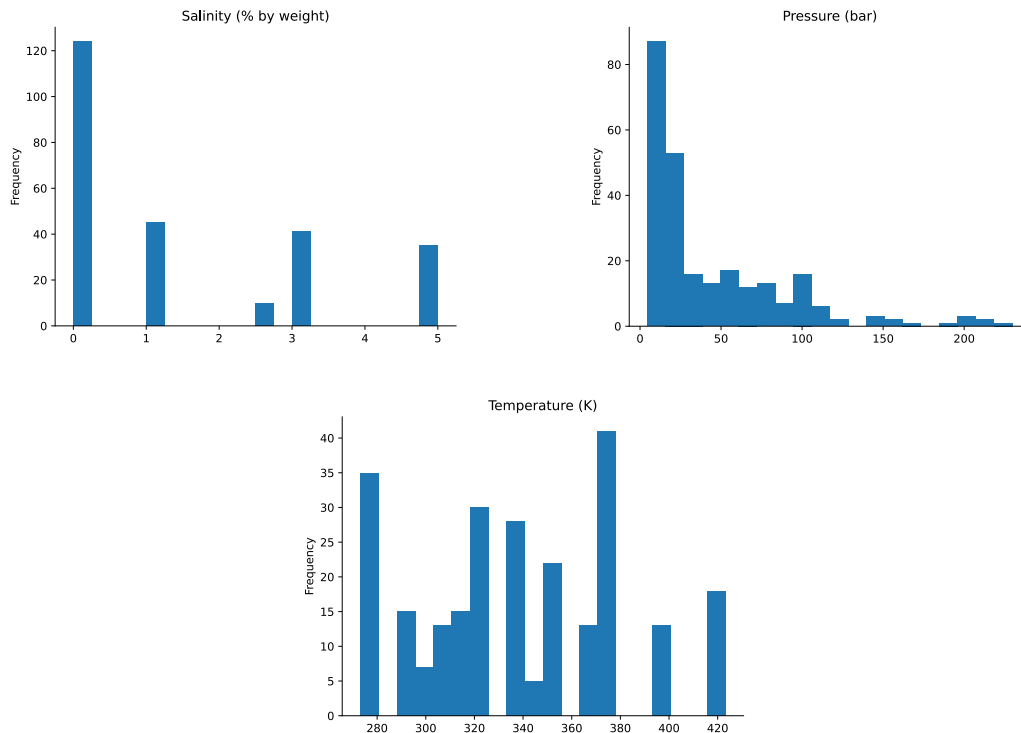


Figure 2: Histograms of variables.

As observed in Figure 3b, the target variable (i.e., hydrogen solubility) showed an asymmetric distribution with many low values and few high values. Additionally, the boxplot also highlighted the presence of several outliers for solubility values above ~ 0.5 mole fraction. The interquartile range (IQR) spanned from approximately 0 to 0.3, indicating that 50% of the data was concentrated within this interval, which can be visualized in Figure 3b.

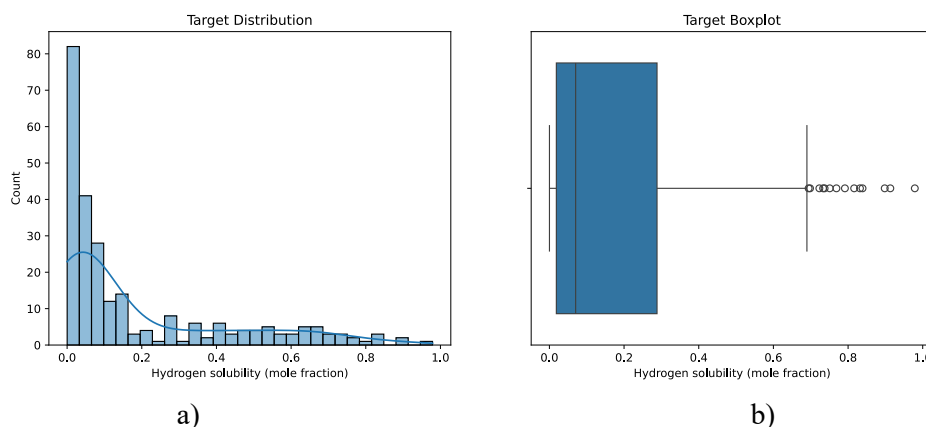


Figure 3: Analysis of the target variable.

The presence of skewness suggested that preprocessing steps might be necessary to improve model performance.

All analyses performed in this study were conducted using Python version 3.11.12.

2.2. Data preprocessing

Data preprocessing was performed as an essential step for maintaining the integrity and reliability of the dataset, thereby establishing a usable format for ML analyses [Zhou et al. 2017].

2.2.1. Feature engineering

Feature engineering was applied to create new attributes from the original variables, enabling the models to capture non-linear relationships and domain-specific interactions that improve predictive performance [Abbasi et al. 2025]. To this end, we manually derived features based on domain knowledge and relevant literature to ensure interpretability and relevance. Specifically, we created interaction, polynomial features, and logarithmic transformations to account for non-linear effects, stabilize variances, and reduce heteroscedasticity. Additionally, we included the pressure and temperature ratio to reflect the thermodynamic relationship between these variables. The additional features and their characteristics are presented in Table 2.

Table 2: Description of the new variables.

Feature	Formula	Purpose	Physical/Statistical Justification
P_T_ratio	P / T	Create a ratio between pressure and temperature.	Reflects thermodynamic equilibrium and hydrogen stability [Bachand et al. 2024].
logP	$\log(1 + P)$	Reduce the scale of pressure variation.	Reduces outlier impact and improves model performance [Solatpour et al. 2024].
P ²	P ²	Capture non-linear effects of pressure	Accounts for non-linear pressure-property relationships [Tiwari et al. 2025].
S_P	$S \times P$	Interaction between salinity and pressure	Captures pressure–salinity interaction in reservoirs [Medina et al. 2024].
T_S	$T \times S$	Interaction between temperature and salinity	Captures temperature–salinity interaction in geological settings [Janjua et al. 2024].
T ²	T ²	Capture non-linear effects of temperature	Models non-linear temperature effects on properties [Tiwari et al. 2025].
logT	$\log(1 + T)$	Smooth the temperature scale	Reduces thermal variation for better modeling [Solatpour et al. 2024].
S ²	S ²	Capture non-linear effects of salinity	Models non-linear salinity effects on hydrogen behavior [Feng et al. 2024].

2.2.2. Data splitting

The dataset was partitioned into an 80:20 ratio, with 0.8 for training and 0.2 for testing, to evaluate model performance on unseen data.

2.2.3. Target transformation

Target transformation is essential when the target variable exhibits strong skewness or deviates significantly from a normal distribution, which can negatively impact the performance of many regression algorithms [Parmo and Wardhana 2024]. By transforming the target variable, we aim to stabilize variance, reduce the influence of outliers, and improve the linearity and homoscedasticity assumptions in predictive models.

In this study, we applied the Yeo-Johnson transformation using *PowerTransformer* from scikit-learn, which can handle both positive and negative values, making it suitable for our dataset. Figure 4 presents the target distribution before and after the transformation.

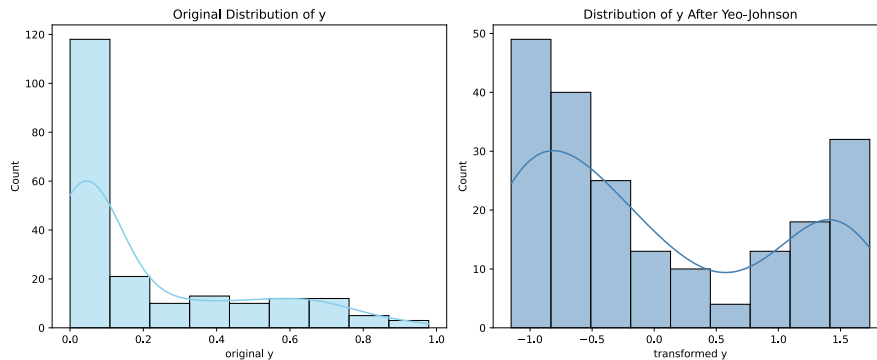


Figure 4: Target distribution before and after transformation.

2.2.4. Standardization

Standardization is critical when working with ML models sensitive to the input features' scale, such as Support Vector Machines. Features with vastly different scales can dominate others, leading to biased learning [Silva et al. 2024a]. To ensure all variables contributed equally during model training and to speed up convergence, we standardized the independent variables using *StandardScaler* from scikit-learn, transforming the features to have a mean of zero and a standard deviation of one.

2.2.5. Feature selection

In this study, feature selection was performed using an automatic stepwise forward selection approach. This iterative process evaluates variables one by one, adding those that contribute most to improving the model's performance while excluding those with minimal or negative impact. The procedure continues until no further improvement is achieved with the addition of new variables [Wang et al. 2022].

The stepwise method identified an optimal subset of predictors, effectively removing redundant or less informative variables. This can be useful to reduce the risk of multicollinearity and simplify the model, enhancing its interpretability and generalizability [Chan et al. 2022]. The final set of selected features consists of 8 features (T_S, logP, S_P, S², Salinity, logT, and Temperature).

2.3. Predictive models

To predict hydrogen solubility, we implemented a range of supervised learning models capable of capturing both linear and non-linear relationships among the selected features.

2.3.1. Models

ML predictive algorithms analyze large datasets to identify patterns that can forecast future events. Their ability to generate significant findings makes them great tools for data-driven strategies across various industries [Maior et al. 2020]. In this study, we applied eight distinct models, summarized in Table 3.

Table 3: Description of the predictive models.

Model	Description
CatBoost	A gradient boosting algorithm specifically optimized for categorical features, effectively reducing overfitting and improving training efficiency.

XGBoost	An efficient gradient boosting model that leverages parallel processing and regularization techniques to achieve high predictive accuracy.
GradientBoost	A sequential model that constructs weak learners (e.g., decision trees) to minimize errors from previous iterations, enhancing predictive performance.
LightGBM	A gradient boosting framework that employs histogram-based learning to significantly accelerate training on large-scale datasets.
Random Forest	An ensemble learning technique based on multiple decision trees, where predictions are aggregated to reduce variance and increase robustness.
Support Vector Regressor	A regression model that utilizes a defined margin of tolerance (i.e., epsilon) to fit data within a specified boundary, ensuring accurate prediction.
Linear Regression	A method that models linear relationships by fitting a straight line to the data, minimizing the sum of squared residuals.
Ridge Regression	A variant of linear regression incorporating L2 regularization to prevent overfitting by penalizing large coefficients.

2.3.2. Hyperparameters optimization

Hyperparameter optimization focuses on enhancing the performance of machine learning models by identifying the most effective configuration of hyperparameters [Silva et al. 2024b]. GridSearchCV is a widely used method that systematically explores a defined search space of hyperparameter values, evaluating each combination through cross-validation to determine the optimal set for a given model [Shams et al. 2023].

3. Results

The study explored different predictive algorithms, such as CatBoost, XGBoost, Gradient Boosting, LightGBM, Linear Regression, Ridge Regression, and Support Vector Regressor, to assess and compare their performance in predicting hydrogen solubility.

The metrics utilized in this study were R^2 , root-mean-squared error (RMSE), and mean absolute error (MAE). These metrics can be calculated using Equations 1–3.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

Where y_i are the actual values, $\hat{y}_i$ are the predicted values, $\bar{y}$ is the mean of the actual values, and n is the total number of observations.

The results indicate that CatBoost achieved the most consistent and superior performance across training, testing, and overall datasets among the tested models. Following the SVR, it reported the second highest R^2 score for the test set, suggesting excellent generalization ability, with low RMSE and MAE, indicating minimal prediction errors. The overall performance (i.e., $R^2 = 0.9939$) reinforces its robustness and accuracy. XGBoost also performed well on the training set, though the slight drop in test performance suggests potential overfitting. Nevertheless, its overall R^2 of 0.9919 and very low MAE of 0.0041 across all data points confirm high predictive capability, as presented in Figure 5. For RMSE and MAE, darker shades represent lower (better) error values,

while for R^2 , darker shades indicate higher (better) scores. This visual mapping helps clearly distinguish models with superior performance.

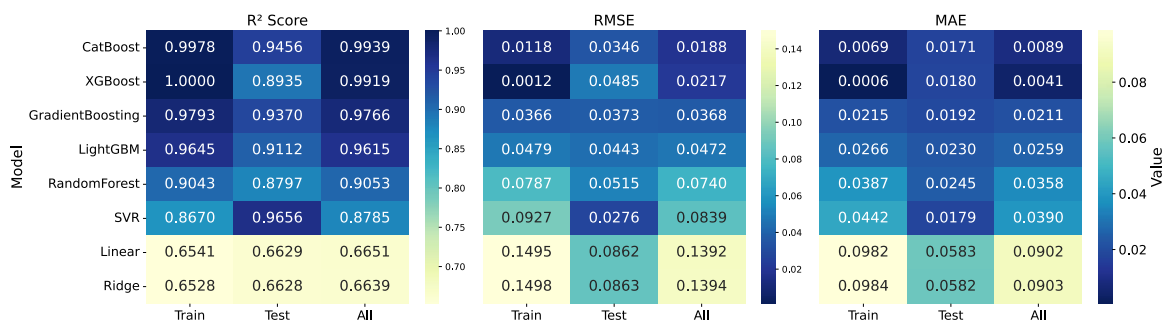


Figure 5: Heatmaps of R^2 , RMSE, and MAE for each model.

Gradient Boosting and LightGBM followed closely, both demonstrating strong generalization with test R^2 values above 0.91. Their slightly higher RMSE and MAE compared to CatBoost and XGBoost suggest modestly reduced precision but still effective performance. Random Forest and SVR delivered moderate performance, with Random Forest showing balanced but lower R^2 scores (i.e., test $R^2 = 0.8797$), and SVR achieving a high test R^2 of 0.9656 despite underperforming on the training set, which may indicate underfitting or sensitivity to parameter tuning. In contrast, Linear Regression and Ridge Regression showed significantly weaker performance, with R^2 scores around 0.66 and higher RMSE and MAE values. This suggests that simple linear models may not adequately capture the complexity and non-linearity present in the data.

The analysis of the CatBoost performance for predicting hydrogen solubility in saline solutions reveals important insights into its behavior. In Figure 6a, the residual distribution is centered around zero, which is a desirable characteristic, as it suggests the model does not exhibit a strong systematic bias, and it does not consistently overpredict or underpredict the values. This balance in the residuals indicates that, on average, the model's predictions are close to the true values, contributing to its overall reliability.

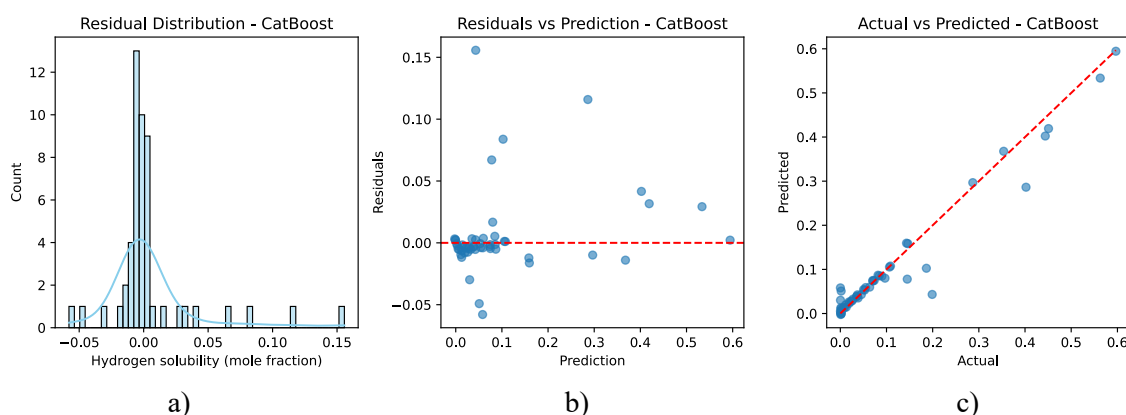


Figure 6: Residuals analysis from CatBoost.

Figure 6b shows the residuals plotted against the predicted values. Here, we observe that the residuals are tightly clustered around zero for lower predicted values, indicating that the model performs well in this range. However, as the predicted values increase, the residuals become more dispersed, with some noticeable positive outliers. This pattern suggests that the model has difficulty

accurately predicting higher values of hydrogen solubility, possibly due to greater complexity in the data or fewer training examples in that range.

In Figure 6c, which compares predicted and actual values, most points align closely with the diagonal red line, confirming a generally good fit. However, there is a growing deviation from the line for higher real values, indicating that the model tends to underestimate the true hydrogen solubility at the upper end of the scale. The clustering of points below the identity line in the higher value range visually confirms this underestimation.

A comparison between the results obtained in this study and those reported in the study of [Vo Thanh et al. 2024] reveals significant differences in the predictive performance of similar ML models, especially in terms of generalization capability on testing data. This analysis can be seen in Table 4.

Table 4: Comparative results analysis from test set.

Model	Metric	This study	Reference study
XGBoost	R ²	0.8935	0.6980
	RMSE	0.0485	0.1358
	MAE	0.0180	0.0432
GradientBoosting	R ²	0.9370	0.9765
	RMSE	0.0373	0.0378
	MAE	0.0192	0.0255
Random Forest	R ²	0.8797	0.6430
	RMSE	0.0515	0.1113
	MAE	0.0245	0.0377

In the reference study, XGBoost achieved near-perfect performance on the training set with $R^2 = 0.9999$, but its performance dropped substantially on the test set with $R^2 = 0.6980$, suggesting overfitting. In the current study, the XGBoost model behaved similarly, with R^2 scores of 1.0000 for the training set and 0.8935 for the test set, reflecting a more robust training process. Gradient Boosting in the reference study outperformed all other models on the test set, presenting $R^2 = 0.9765$ and $RMSE = 0.0378$, slightly exceeding the performance observed in the current study with $R^2 = 0.9370$ and $RMSE = 0.0373$. Yet the differences are minimal, and both studies emphasize the strong predictive capacity of Gradient Boosting. Random Forest, however, showed considerably weaker generalization in the reference study (i.e., $R^2 = 0.6430$) compared to the current one (i.e., $R^2 = 0.8797$), again suggesting that the model tuning and preprocessing pipeline adopted in this work were more effective in preventing overfitting and enhancing performance.

4. Conclusion

This study explored the application of ML models to predict hydrogen solubility in saline aqueous solutions, aiming to support underground storage strategies. Among the models tested, CatBoost achieved the best overall performance, with a high coefficient of determination of $R^2 = 0.9939$, a low root mean square error (i.e., $RMSE = 0.0188$), and a mean absolute error (i.e., $MAE = 0.0089$), indicating strong predictive accuracy and generalization ability. XGBoost also performed exceptionally well in the training phase, reaching a perfect R^2 of 1; however, its lower performance in the testing phase suggests potential overfitting to the training data. The Support Vector Regressor demonstrated consistent results, showing strong generalization on the test set, while linear models such as Linear Regression and Ridge Regression showed significantly lower performance, with R^2 values close to 0.66, reinforcing the need for non-linear models in this context.

Despite these promising results, some limitations must be considered. The dataset size was relatively small, which may restrict the models' ability to generalize to new or unseen data.

This limitation is particularly relevant when dealing with complex chemical systems, where the influence of various parameters can be highly non-linear and interdependent. Moreover, the study did not include a computational cost analysis, which is essential when considering real-world implementation, especially for large-scale or real-time prediction systems.

For future research, expanding the dataset to include a broader range of salinity conditions and temperature regimes could enhance the robustness of the models. Incorporating data from critical solubility regions would be particularly valuable for improving accuracy where precision is most needed. It is also advisable to explore alternative feature selection strategies, beyond the stepwise approach, including methods based on feature importance from ensemble models. Additionally, investigating the use of neural networks or hybrid models may provide further improvements by capturing complex non-linear relationships that are not easily modeled by traditional algorithms. Lastly, assessing the computational efficiency of each model would offer practical insights into selecting the most viable algorithm for application in operational hydrogen storage systems.

Acknowledgments

The authors thank the Brazilian research funding agencies Human Resources Program (PRH 38.1) entitled “Risk Analysis and Environmental Modeling in the Exploration, Development and Production of Oil and Gas”, financed by ‘Agência Nacional de Petróleo (ANP)’ and managed by FAPESP, the ‘Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq): 402761/2023-5’, the ‘Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE): APQ-0504-3.08/25’, and ‘Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)’ - Finance Code 001 for the financial support through research grants.

References

- Abbasi, M., Váz, P., Silva, J., & Martins, P. (2025). Machine Learning Approaches for Predicting Maize Biomass Yield: Leveraging Feature Engineering and Comprehensive Data Integration. *Sustainability*, 17(1), 256. <https://doi.org/10.3390/su17010256>
- Bachand, A., Doyon, B., & Raymond, J. (2024). Thermo-physical numerical model for hydrogen storage in underground tanks and caverns. *International Journal of Hydrogen Energy*, 66, 66–80. <https://doi.org/10.1016/j.ijhydene.2024.03.246>
- Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, 10(8), 1283. <https://doi.org/10.3390/math10081283>
- Dehghani, M. R., Nikravesh, H., Aghel, M., Kafi, M., Kazemzadeh, Y., & Ranjbar, A. (2024). Estimation of hydrogen solubility in aqueous solutions using machine learning techniques for hydrogen storage in deep saline aquifers. *Scientific Reports*, 14(1), 25890. <https://doi.org/10.1038/s41598-024-76850-8>
- Feng, X., Liu, J., Shi, J., Hu, P., Zhang, T., & Sun, S. (2024). Phase equilibrium, thermodynamics, hydrogen-induced effects and the interplay mechanisms in underground hydrogen storage. *Computational Energy Science*, 1(1), 46–64. <https://doi.org/10.46690/compes.2024.01.05>
- Gbadamosi, A. O., Muhammed, N. S., Patil, S., Al Shehri, D., Haq, B., Epelle, E. I., Mahmoud, M., & Kamal, M. S. (2023). Underground hydrogen storage: A critical assessment of fluid-fluid

and fluid-rock interactions. *Journal of Energy Storage*, 72, 108473. <https://doi.org/10.1016/j.est.2023.108473>

Hassan, Q., Viktor, P., J. Al-Musawi, T., Mahmood Ali, B., Algburi, S., Alzoubi, H. M., Khudhair Al-Jiboory, A., Zuhair Sameen, A., Salman, H. M., & Jaszczur, M. (2024). The renewable energy role in the global energy Transformations. *Renewable Energy Focus*, 48, 100545. <https://doi.org/10.1016/j.ref.2024.100545>

Janjua, A. N., Ali, M., Murtaza, M., Patil, S., & Kamal, M. S. (2024). Effects of salinity, temperature, and pressure on H₂–brine interfacial tension: Implications for underground hydrogen storage. *Journal of Energy Storage*, 95, 112510. <https://doi.org/10.1016/j.est.2024.112510>

Keith, J. A., Vassilev-Galindo, V., Cheng, B., Chmiela, S., Gastegger, M., Müller, K.-R., & Tkatchenko, A. (2021). Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chemical Reviews*, 121(16), 9816–9872. <https://doi.org/10.1021/acs.chemrev.1c00107>

Li, D., Beyer, C., & Bauer, S. (2018). A unified phase equilibrium model for hydrogen solubility and solution density. *International Journal of Hydrogen Energy*, 43(1), 512–529. <https://doi.org/10.1016/j.ijhydene.2017.07.228>

Maior, C. B. S., Moura, M. J. das C., Santana, J. M. M., & Lins, I. D. (2020). Real-time classification for autonomous drowsiness detection using eye aspect ratio. *Expert Systems with Applications*, 158, 113505. <https://doi.org/10.1016/j.eswa.2020.113505>

Maior, C. B. S., Silva, S. P., Lins, I. D., Lisa Gomes, A., & Moura, M. C. (2025). Machine Learning for Long COVID Inference Based on the Acute Phase: A Case Study in Healthcare Professionals. *IEEE Access*, 13, 54019–54027. <https://doi.org/10.1109/ACCESS.2025.3553783>

Medina, O. E., Gallego, J. F., Moncayo-Riascos, I., Lysyy, M., Benjumea, P. N., Cortés, F. B., & Franco, C. A. (2024). Salinity influence on underground hydrogen storage: Insights from molecular dynamics and pore-scale analysis. *International Journal of Hydrogen Energy*, 60, 959–975. <https://doi.org/10.1016/j.ijhydene.2024.02.073>

Mwakipunda, G. C., Komba, N. A., Kouassi, A. K. F., Ayimadu, E. T., Mgimba, M. M., Ngata, M. R., & Yu, L. (2024). Prediction of hydrogen solubility in aqueous solution using modified mixed effects random forest based on particle swarm optimization for underground hydrogen storage. *International Journal of Hydrogen Energy*, 87, 373–388. <https://doi.org/10.1016/j.ijhydene.2024.09.054>

Parmo, P., & Wardhana, I. (2024). Computational Engineering and Physical Modeling Investigating the Influence of Cement and Additive Normalization on Concrete Compressive Strength: A Statistical Analysis ARTICLE INFO ABSTRACT. P. Parmo; I. Wardhana/ *Computational Engineering and Physical Modeling*, 7(4), 22–44. <https://doi.org/10.22115/cepm.2024.484143.1335>

Ramesh Kumar, K., Honorio, H., Chandra, D., Lesueur, M., & Hajibeygi, H. (2023). Comprehensive review of geomechanics of underground hydrogen storage in depleted reservoirs and salt caverns. *Journal of Energy Storage*, 73, 108912. <https://doi.org/10.1016/j.est.2023.108912>

- Semedo, A., Garcia, J., & Brito, M. (2025). Cryogenics in Renewable Energy Storage: A Review of Technologies. *Energies*, 18(6), 1543. <https://doi.org/10.3390/en18061543>
- Shams, M. Y., Elshewey, A. M., El-kenawy, E.-S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2023). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, 83(12), 35307–35334. <https://doi.org/10.1007/s11042-023-16737-4>
- Silva, S. P. da, Maior, C. B. S., Lins, I. D., & Moura, M. J. D. C. (2024a). *Machine learning algorithms for predicting hydrogen embrittlement in steel pipelines: a focus on hydrogen gas transportation*. 4831–4845. <https://doi.org/10.48072/2525-7579.roge.2024.3364>
- Silva, S. P. da, Maior, C. S., Lins, I. D., & Moura, M. das C. (2024b). *Exploring Data-Driven Solutions for Hydrogen Safety: Predictive Modeling of Explosion Thresholds*. www.portalabpg.org.br
- Solatpour, R., Babak, P., & Kantzas, A. (2024). Log-exponential transformation function for interpreting NMR relaxation measurements of hydrocarbon in organic porous media for enhancing absolute adsorption estimation. *Chemical Engineering Science*, 286, 119607. <https://doi.org/10.1016/j.ces.2023.119607>
- Tiwari, P., Naskar, S., & Mukhopadhyay, T. (2025). Non-linear functionally graded metamaterials for hydrogen storage and enhanced sustainability under extreme environments. *Thin-Walled Structures*, 210, 112901. <https://doi.org/10.1016/j.tws.2024.112901>
- Vo Thanh, H., Zhang, H., Dai, Z., Zhang, T., Tangparitkul, S., & Min, B. (2024). Data-driven machine learning models for the prediction of hydrogen solubility in aqueous systems of varying salinity: Implications for underground hydrogen storage. *International Journal of Hydrogen Energy*, 55, 1422–1433. <https://doi.org/10.1016/j.ijhydene.2023.12.131>
- Wang, C., Xu, S., Liu, J., Yang, J., & Liu, C. (2022). Building an improved artificial neural network model based on deeply optimizing the input variables to enhance rutting prediction. *Construction and Building Materials*, 348, 128658. <https://doi.org/10.1016/j.conbuildmat.2022.128658>
- Zhang, L., Jia, C., Bai, F., Wang, W., An, S., Zhao, K., Li, Z., Li, J., & Sun, H. (2024). A comprehensive review of the promising clean energy carrier: Hydrogen production, transportation, storage, and utilization (HPTSU) technologies. *Fuel*, 355, 129455. <https://doi.org/10.1016/j.fuel.2023.129455>
- Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237, 350–361. <https://doi.org/10.1016/j.neucom.2017.01.026>
- Zhou, Z., Nourani, P., Karimi, M., Kamrani, E., & Anqi, A. E. (2022). Relying on machine learning methods for predicting hydrogen solubility in different alcoholic solvents. *International Journal of Hydrogen Energy*, 47(9), 5817–5827. <https://doi.org/10.1016/j.ijhydene.2021.11.121>