China Geology 9 (2026) 519−534

China Geology

Journal homepage: http://chinageology.cgs.cn
https://www.sciencedirect.com/journal/china-geology

Machine learning supervised algorithms for gas hydrate identification and saturation estimation in marine reservoirs using well log data: A case study of NGHP-01-19B

Yi-fan Wu[a, b,*], Zheng Su[a, b,*], Takeshi Tsuji[c], Dai-dai Wu[a, b], Guang-rong Jin[a, b], Chao Yang[a, b], Chuang-ji Feng[a, b], Neng-you Wu[d, e]

[a] *School of Energy Science and Engineering*, *University of Science and Technology of China*, *Guangzhou* 510640, *China*
[b] *Guangzhou Institute of Energy Conversion*, *Chinese Academy of Sciences*, *Guangzhou* 510640, *China*
[c] *Department of Systems Innovation*, *School of Engineering*, *The University of Tokyo*, *Tokyo* 113-8656, *Japan*
[d] *Key Laboratory of Gas Hydrate*, *Ministry of Natural Resources*, *Qingdao Institute of Marine Geology*, *China Geological Survey*, *Qingdao* 266237, *China*
[e] *Laoshan Laboratory*, *Qingdao* 266200, *China*

ARTICLE INFO

Article history:
Received 17 April 2025
Received in revised form 30 June 2025
Accepted 15 July 2025
Available online 28 September 2025

Keywords:
Gas hydrate
Machine learning algorithm
Classification
Regression
Well log data
Archie's law
Marine geohazards
Global carbon cycle

ABSTRACT

Gas hydrates are increasingly recognized as a significant unconventional energy resource and a key factor in marine geohazards and the global carbon cycle. However, accurately identifying and quantifying hydrate-bearing formations remains challenging due to complex geophysical signatures and heterogeneous distribution. This study evaluates twelve supervised machine learning (ML) algorithms for two key tasks: Classification of hydrate-bearing layers and regression-based estimation of hydrate saturation, using well log and pore-water geochemical data from Site NGHP-01-19B. Two physically independent labeling frameworks are employed: One based on Archie's law using resistivity (1350 samples, 29% hydrate-bearing), and another based on a three-phase velocity model (890 samples, 25% hydrate-bearing). A diverse set of models, including tree-based ensembles (Decision Tree, Random Forest, GBDT, XGBoost, LightGBM, CatBoost, Bagging, AdaBoost), kernel methods (SVM, SVR), instance-based learning (KNN), neural networks (MLP), and Gaussian Process models (GPR, GPC), are systematically compared using cross-validation and grid search. Ensemble methods consistently performed best in classification, with AdaBoost and GBDT, achieving test accuracies above 0.94 (Archie) and 0.98 (velocity-based). For regression, GPR delivered the most accurate hydrate saturation estimates ($R^2 > 0.99$), while GBDT and Random Forest provided a strong balance of accuracy and computational efficiency. Notably, depth below seafloor (TDEP), though not a direct geophysical input, significantly enhanced model performance by acting as a proxy for stratigraphic and thermodynamic conditions. Group-based validation confirmed that random-sample splitting overestimates performance due to depth-wise autocorrelation, highlighting the importance of geologically informed model assessment. Overall, the consistent performance of ML models across both labeling schemes and input feature sets underscores their robustness and transferability, supporting their use as a reliable toolset for offshore gas hydrate reservoir characterization.

©2026 China Geology Editorial Office.

1. Introduction

Gas hydrates are crystalline compounds in which gas molecules are trapped within a solid water lattice, forming a cage-like structure (Sloan ED, 1991). Gas hydrates, with reserves estimated to exceed those of conventional natural gas, represent a promising alternative energy source (Collett TS, 2002). Beyond their economic value, they play a crucial role in the global carbon cycle (Dickens GR, 2003), greenhouse gas storage (Milkov AV and Sassen R, 2003), and marine geohazards (Sultan N et al., 2004). These factors have driven extensive research on gas hydrate exploration and characterization (Milkov AV, 2004; Su Z et al., 2012, 2013), with a particular focus on improving resource assessment and identification accuracy (Jana S et al., 2017; Yuan HM et al., 2021).

Driven by rapid advances in artificial intelligence,

First author: *E-mail address*: wyf0924@mail.ustc.edu.cn (Yi-fan Wu).

* Corresponding author: *E-mail address*: wyf0924@mail.ustc.edu.cn (Yi-fan Wu); suzheng@ms.giec.ac.cn (Zheng Su).

Literary editor: Li-qiong Jia
doi:10.31035/cg2025078
2096-5192/© 2026 China Geology Editorial Office.

Copyright © 2026 Editorial Office of China Geology. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd.
This is an open access article under the CC BY-NC-ND License (http://creativecommons.org/licenses/by-nc-nd/4.0/).

machine learning (ML) has become an increasingly powerful tool in geosciences, enabling applications in reservoir evaluation, lithology identification, and gas hydrate exploration (Chong L et al., 2024). Recent studies have demonstrated the potential of ML techniques in hydrate occurrence classification, saturation prediction, and formation condition modeling, often surpassing traditional physics-based approaches. For instance, machine learning algorithms such as Random Forest (RF), AdaBoost, and Support Vector Regression (SVR) have been applied in hydrate-bearing layer identification, exhibiting varying performance depending on geological context (Tian DM et al., 2023). Similarly, ML-based regression models, particularly Artificial Neural Networks (ANN), have been used to predict gas hydrate saturation with notable accuracy improvements over conventional methods (Singh H et al., 2021; Chong L et al., 2022).

Beyond classification and saturation estimation, ML has found applications in deep-sea porosity estimation, seismic structure analysis, and hydrate formation modeling (Xu HF et al., 2021; Zhu LQ et al., 2022). Deep learning models, such as Convolutional Neural Networks (CNNs), have proven effective in seismic interpretation, automating the detection of slump units and fault structures (Ahmad AB and Tsuji T, 2021). In reservoir classification, ML clustering methods like K-means and ensemble tree-based models have shown strong performance in delineating gas hydrate reservoirs based on heterogeneity patterns (Zhu LQ et al., 2023). Collectively, these studies highlight the growing role of machine learning across multiple aspects of hydrate-related research, including reservoir classification, porosity estimation, saturation prediction, and subsurface structure interpretation.

Building on recent progress, this study systematically evaluates a range of supervised learning algorithms for two key tasks: Gas hydrate occurrence classification (classification task) and gas hydrate saturation estimation (regression task). While prior research has demonstrated ML's effectiveness in specific applications, a comprehensive comparison of mainstream ML models across both predictive tasks remains limited. To address this gap, a diverse set of supervised algorithms is employed, including Decision Tree, Bagging, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT, XGBoost, LightGBM, CatBoost), AdaBoost and Multilayer Perceptron (MLP) for classification, alongside Gaussian Process models (GPC, GPR) for saturation prediction.

To ensure robust model evaluation, feature engineering, hyperparameter tuning via grid search, and cross-validation are incorporated. Model performance is assessed using standard classification metrics (Accuracy, Precision, Recall, F1–score, AUC) and regression evaluation metrics (MSE, RMSE, R^2). Through a systematic comparison, this study aims to develop an efficient and accurate ML-based approach for hydrate reservoir characterization, ultimately contributing to improved predictive modeling and decision-making in gas hydrate exploration.

To this end, this study evaluates twelve supervised learning algorithms across both classification and regression tasks, providing a broad performance landscape for hydrate detection and saturation estimation. A physically decoupled framework is designed, in which two independent labeling strategies—resistivity-based and velocity-based—are paired with mutually exclusive input features to minimize label leakage and enable cross-label evaluation of model robustness and transferability. A geologically informed, group-based validation scheme is further adopted to assess generalization under realistic deployment conditions and to address the performance overestimation caused by depth-wise autocorrelation in random sampling. Finally, the trade-off between predictive accuracy and computational efficiency is examined to provide practical guidance for selecting scalable algorithms in offshore hydrate exploration workflows.

2. Geological setting and data

The Mahanadi Basin is a sedimentary basin on India's eastern coast, extending from Jagannathpur to Chilka Lake and further offshore into the Bay of Bengal (Mohapatra P, 2006). Formed during the Jurassic rifting of Gondwanaland (Sastri VV et al., 1981), it covers about 2.6×10^5 km^2, with the offshore portion spanning about 8×10^4 km^2 and water depths ranging from 100 m to 2500 m (Dangwal V et al., 2008). The basin has accumulated over 8 km of Upper Cretaceous to recent sediments, primarily deposited by the Ganga–Brahmaputra deltaic system during the Mio–Pliocene (Bharali B, 1991). Tectonically, the basin extends from the Damra–Wheeler Lineament in the north to the 85° East Ridge in the south (Mohapatra P, 2006), which separates it from the Krishna–Godavari (KG) Basin, a significant hydrocarbon province (Jagannathan CR et al., 1984; Fuloria RC et al., 1992; Bastia R, 2006). Seismic imaging has identified a submarine ridge south of Chilka Lake, which played a crucial role in sediment confinement and gas hydrate accumulation (Bastia R et al., 2010).

This study analyzes well log data from Site NGHP-01-19B, drilled during India's National Gas Hydrate Program Expedition 01 (NGHP-01). The NGHP-01 conducted scientific drilling, coring, and geophysical logging to evaluate gas hydrate occurrence along India's continental margins (Collett TS et al., 2014). Located in the Mahanadi Basin, Site NGHP-01-19B serves as a key exploration site for gas hydrates (Fig. 1). The dataset includes resistivity, sonic velocity (V_P, V_S), bulk density, and porosity logs.

Wire-line logging at Site NGHP-01-19 provided key geophysical measurements for assessing gas hydrate occurrence, including electrical resistivity, sonic velocity (V_P, V_S), neutron porosity, and bulk density (Collett TS et al., 2014). Data quality is lower in the upper 80 mbsf (meters below seafloor) due to borehole conditions but higher with depth, enabling more reliable analysis at greater depths.

Sonic velocity logs, acquired using the Dipole Sonic Imager (DSI), measure P-wave (V_P) and S-wave (V_S)

velocities with a vertical resolution of about 1 m. The velocity profile (Fig. 2) shows a general increasing trend with depth, peaking just above the bottom-simulating reflector (BSR) at about 205 mbsf. Below the BSR, a sharp drop in P-wave velocity (about 1.53 km/s) suggests the presence of free gas, followed by a gradual increase to about 1.63 km/s (Collett TS et al., 2008).

Resistivity logs, obtained using Schlumberger wire-line tools, have a vertical resolution of about 2 m and an accuracy of ±10% (Shankar U and Riedel M, 2014). Among all geophysical logs, resistivity is the most sensitive to gas hydrate presence, as gas hydrate formation reduces porosity and electrical conductivity, leading to elevated resistivity compared to water-saturated sediments (Yuan T et al., 1996; Collett TS and Ladd J, 2000; Wang XJ et al., 2011). At NGHP-01-19, background resistivity values in hydrate-free sediments range from 0.7 Ω·m to 0.9 Ω·m, whereas intervals 175–190 mbsf exhibit elevated resistivity, indicating gas hydrate occurrence (Fig. 2). The base of gas hydrate stability zone (BGHSZ) is marked by a distinct resistivity drop near 205 mbsf, aligning with seismic BSR interpretations.

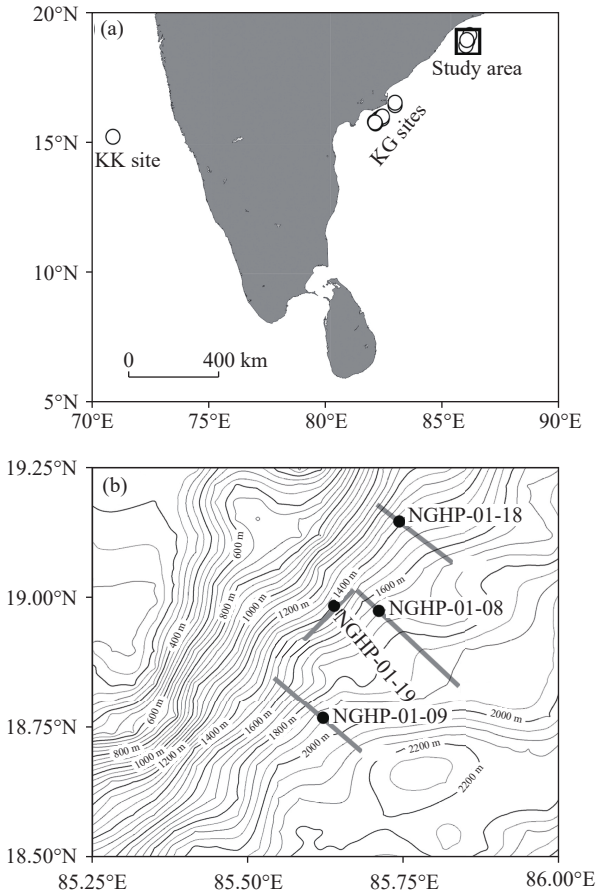

Fig. 1. a–Map of study area showing the locations of the drilling sites NGHP-01-19 in the Mahanadi Basin drilled during NGHP Expedition-01; b–detail bathymetry of study area with drill site location (black filled circle). The bold solid lines mark the transect of the seismic lines, which are crossing or adjacent to the drilling sites shown (modified from Shankar U and Riedel M, 2014).

To improve data quality, a series of preprocessing steps were applied to the raw logging data. First, all geophysical parameters were depth-aligned to the depth below seafloor (TDEP) scale to correct for discrepancies in tool-specific measurement origins and calibration offsets. Subsequently, anomalous values caused by borehole enlargements, tool malfunctions, or environmental noise were identified and removed. Finally, because the original logging measurements were acquired at varying depth intervals across different tools, linear interpolation was performed to resample all datasets onto a consistent depth grid, ensuring uniformity for subsequent analysis. In addition to geophysical logs, pore-water geochemical measurements, including chloride concentration and salinity, were also incorporated. These values were depth-aligned and interpolated to match the log grid for consistency in subsequent analysis (Fig. 3).

3. Method

3.1. Gas hydrate saturation estimation methods

3.1.1. Archie's law (resistivity-based)

Gas hydrate saturation (S_h) at Site NGHP-01-19B was estimated using Archie's Law (Archie GE, 1942), an empirical relationship that linked electrical resistivity to porosity and water saturation. Originally developed for hydrocarbon reservoirs, Archie's equation has been widely adopted in gas hydrate studies to infer pore fluid saturation based on resistivity measurements (Collett TS and Ladd J, 2000). The key parameter in this approach is the formation

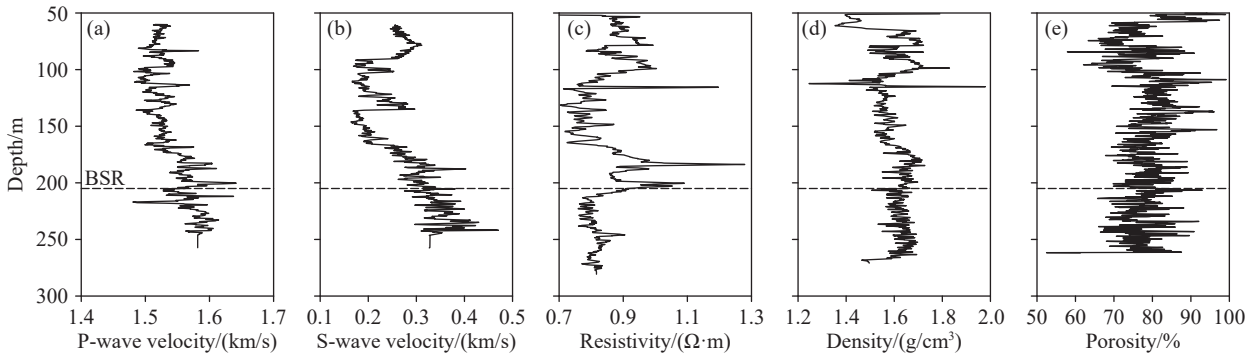

Fig. 2. Well-log data from Site NGHP-01-19. a–P-wave velocity; b–S-wave velocity; c–electrical resistivity; d–bulk density; e–porosity. The dotted horizontal line denotes the location of the seismically inferred bottom simulating reflector (BSR) (modified from Shankar U and Riedel M, 2014).

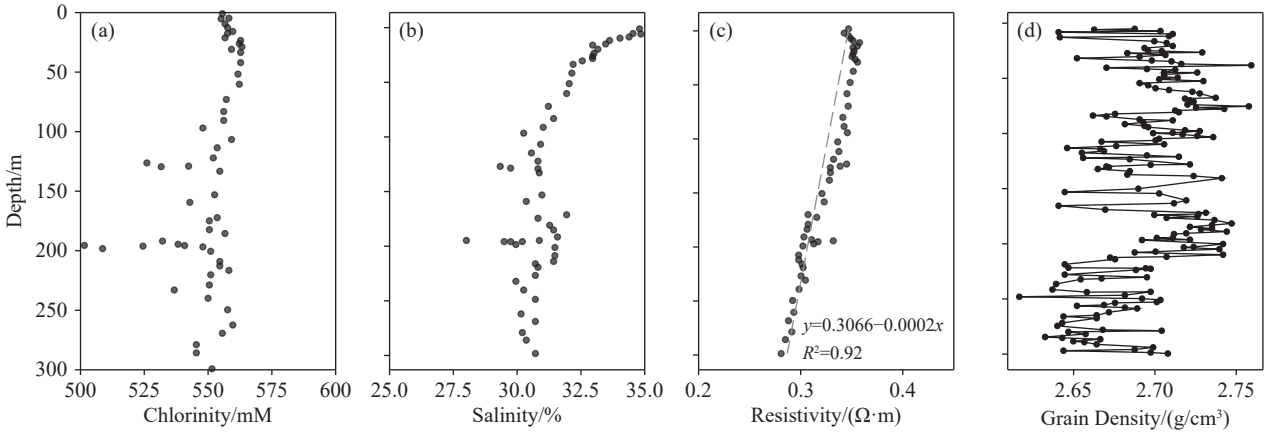

Fig. 3. Petrophysical properties from Site NGHP-01-19B. a–Pore-water chloride concentration profile; b–measured pore-water salinity (S_w) values from sediment cores; c–calculated pore-water resistivity (R_w), with an empirical linear fit (R^2 = 0.92) indicating the background trend; d–grain density profile (modified from Shankar U and Riedel M, 2014).

factor (*FF*), defined as (Eq. 1):

$$FF = \frac{R_o}{R_w} = a \cdot \phi^{-m} \tag{1}$$

where R_o is the resistivity of fully water-saturated sediments, R_w is the pore-water resistivity, ϕ is porosity (derived from bulk density logs), and a and m are empirical Archie parameters.

Pore-water resistivity (R_w) was calculated using the equation of state for seawater (Fofonoff NP, 1985), incorporating hydrostatic pressure, *in-situ* temperature, and core salinity data from NGHP-01-19B (Collett TS et al., 2014). The depth profile of R_w, shown in Fig. 3c (Shankar U and Riedel M, 2014), aligns well with the fitted trend (R^2 = 0.92), supporting its reliability. The Archie parameters (a, m) were obtained through Pickett plots (Collett TS and Ladd J, 2000) by plotting *FF vs*. density porosity for water-saturated sediments (Fig. 4). Low-resistivity zones were used to define the water-saturated baseline, while grain density, derived from core analyses, increases with depth from about 2.65 g/cm^3 at the seafloor to about 2.7 g/cm^3 at the borehole bottom.

Gas hydrate saturation (S_h) was estimated using the following equation (Eq. 2):

$$S_h = 1 - \left(\frac{a \cdot R_w}{\phi^m \cdot R_t}\right)^{\frac{1}{n}} \tag{2}$$

where n is the saturation exponent, which varies with lithology and grain size distribution (Spangenberg E, 2001). While n is not strictly constant, this study adopts a fixed n = 1.94, based on marine sediments (Pearson CF et al., 1983). Although a constant n may underestimate local S_h, site-specific values can be refined by comparing resistivity-based estimates with chlorinity-derived values (Malinverno A et al., 2008).

The S_h–depth profile (Fig. 5a) indicates gas hydrate occupies up to 15% of the pore space in a 25 m-thick zone above the BSR. Below the BSR (about 205 mbsf), S_h drops sharply, likely due to the presence of free gas, as indicated by a marked reduction in acoustic velocity between 205 mbsf and 225 mbsf. These estimates are consistent with results from pressure-core degassing experiments (Collett TS et al., 2015), supporting the validity of the Archie-based approach.

3.1.2. Three-phase weighted equation (velocity-based)

In addition to the resistivity-based method, gas hydrate saturation was also estimated using a velocity-based approach following the formulation of Shankar U and Riedel M (2014). This method combines two classical end-member models for acoustic wave propagation in sediments: The time-average equation (Wyllie MRJ et al., 1958), which is suitable for consolidated formations, and the Wood equation (Wood AB, 1941), which applies to suspensions or highly porous media. Recognizing that neither model alone adequately captures acoustic behavior in unconsolidated marine sediments, a weighted average of the two was adopted to estimate P-wave velocity in hydrate-bearing zones.

The final velocity equation includes an empirical

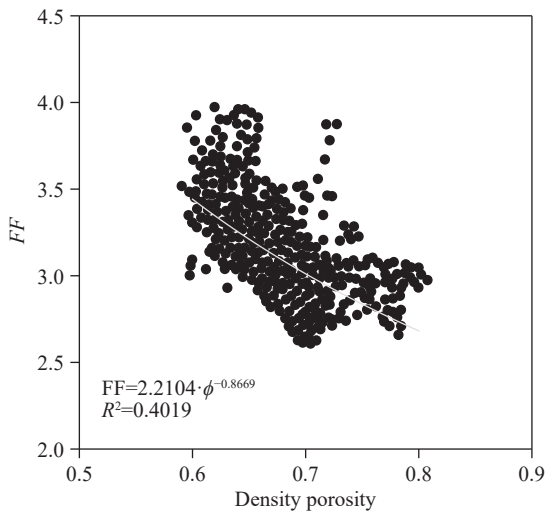

Fig. 4. Pickett plot of formation factor (*FF*) *vs*. density porosity for non-hydrate-bearing sediments and sub-BSR intervals at Site NGHP-01-19B. The exponential best-fit trend (gray line) is used to derive empirical Archie parameters a and m, following the method of Shankar U and Riedel M (2014).

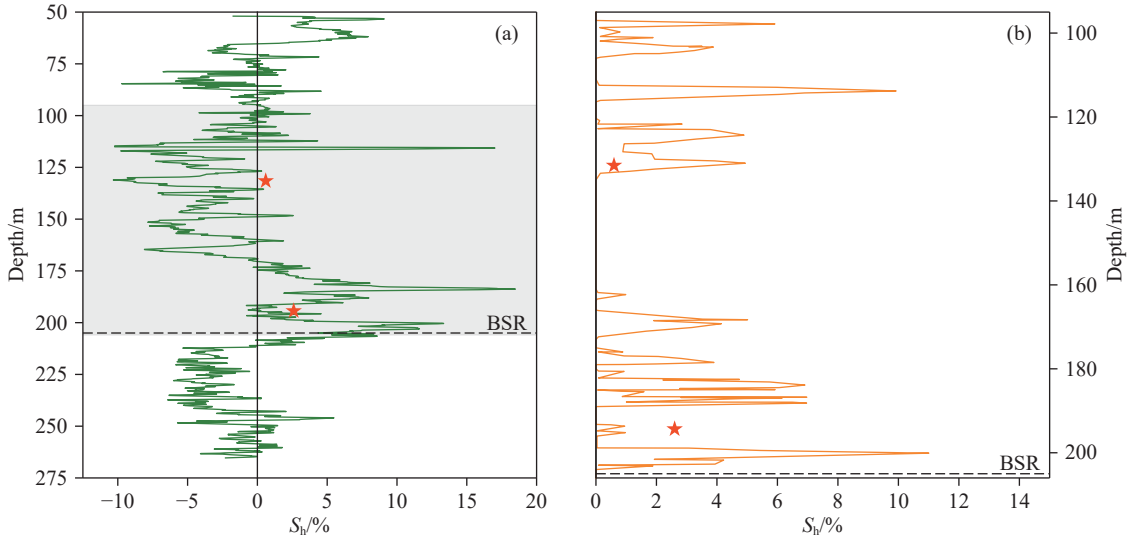

Fig. 5. Gas hydrate saturation (S_h) profiles at Site NGHP-01-19B derived from resistivity- and velocity-based methods. a–S_h estimated using Archie's law with resistivity and density-derived porosity logs; b–S_h estimated using the three-phase weighted equation with acoustic velocity logs. Red pentagram symbols indicate gas hydrate saturation values obtained from pressure-core degassing experiments (Collett TS et al., 2014). Both log-derived estimates follow methods established by Shankar U and Riedel M (2014) and Collett TS et al. (2014). Values below the base of the gas hydrate stability zone (BGHSZ, about 205 mbsf) may reflect the presence of free gas rather than gas hydrate.

weighting factor W, which balances the contributions of the Wyllie and Wood models. Shankar U and Riedel M (2014) determined that a best-fit value of W = 1.27 yields optimal velocity predictions in hydrate-free zones at Site NGHP-01-19B. This approach, often referred to as a three-phase weighted equation, accounts for the combined effects of gas hydrate, pore fluid, and mineral grains on acoustic velocity in unconsolidated sediments.

In this study, the velocity-based gas hydrate saturation profile (Fig. 5b) was adapted from the published results of Shankar U and Riedel M (2014), using their estimated values for Site NGHP-01-19B. These values provide an independent reference to complement the Archie-derived saturation profile and support the training of machine learning regression models.

3.2. Machine learning methodology

To comprehensively assess ML performance in gas hydrate analysis, the authors formulate the problem as two supervised learning tasks. (1) Classification Task: Predict the presence or absence of gas hydrate at a given depth based on well log data (0: No hydrate; 1: Hydrate present). (2) Regression Task: Estimate gas hydrate saturation (S_h) as a continuous variable to enable refined reservoir characterization.

To address these tasks, multiple supervised learning models were evaluated, including Decision Tree (DT), Random Forest (RF), Gradient Boosting Methods (GBDT, XGBoost, LightGBM, CatBoost), Support Vector Machine (SVM, SVR), K-Nearest Neighbors (KNN), Bagging, AdaBoost, MLP, and Gaussian Process models (GPC, GPR). Hyperparameter tuning was performed to optimize model performance and select the most effective algorithm.

3.2.1. Feature selection

The predictive performance of machine learning models depends critically on the choice and physical independence of input features.

In this study, the authors primarily evaluate supervised learning under an Archie-based label framework. To evaluate model robustness and generalization of model predictions across different saturation definitions, an auxiliary configuration using a velocity-derived label (based on the three-phase weighted equation) is also included.

To avoid target leakage and enhance model interpretability, input features are selected to be physically decoupled from the variables used in label derivation under each method.

(i) Archie-based configuration. Input features: V_P (P-wave velocity), V_S (S-wave velocity), and TDEP.

These features are not involved in the computation of Archie-derived S_h, which is based on resistivity, porosity, pore-water chloride concentration, and core-measured salinity (Figs. 2, 3). Thus, the input and target variables remain physically distinct. This configuration includes 1350 depth-aligned samples.

(ii) Three-phase weighted equation configuration. Input features: Formation resistivity, chloride concentration (Cl^-), and core-derived salinity.

The target saturation in this setup is derived from acoustic log data using the three-phase weighted velocity model. Notably, P-wave velocity itself is excluded from the input features to maintain separation from the target variable. This configuration includes 890 matched feature–label pairs.

This mutual exclusivity between input features and label-defining equations in both configurations ensures that the models do not inadvertently learn circular or tautological relationships.

While the Archie-based framework serves as the primary experimental basis for model development and evaluation, the three-phase weighted configuration is introduced as an auxiliary setup to test the robustness and generalizability of the models under alternative saturation definitions.

This two-label design improves the credibility of the results across both classification (hydrate presence) and regression (saturation estimation) tasks, while highlighting the adaptability of the models to distinct geophysical labeling schemes.

All input features are standardized prior to training.

3.2.2. Machine learning algorithms and model training

The dataset was divided into training (80%) and test (20%) subsets, and five-fold cross-validation was employed to evaluate model generalization. Hyperparameter optimization was conducted via grid search (GridSearchCV). All models were implemented using the Scikit-learn library (Pedregosa F et al., 2011; Buitinck L et al., 2013).

Tree-based models, such as DT, RF and GBDT, formed the backbone of this study due to their robustness and compatibility with tabular geophysical data. RF reduces variance through bootstrapped sampling and random feature selection, while GBDT fits residuals sequentially to improve accuracy (Breiman L, 2001; Harris JR and Grunsky EC, 2015; Ma XL et al., 2017; Zhao XB et al., 2022). To further enhance performance and efficiency, the authors also evaluated advanced boosting frameworks such as XGBoost, LightGBM, and CatBoost. These models offer improvements in regularization, handling of missing data, and training speed, making them particularly suited for large-scale applications (Chen TQ and Guestrin C, 2016; Ke GL et al., 2017; Prokhorenkova L et al., 2018; Nguyen H et al., 2019; Hancock JT and Khoshgoftaar TM, 2020).

In contrast, Bagging and AdaBoost were included as classic ensemble methods to benchmark simpler aggregation strategies. Bagging averages multiple base learners trained on bootstrapped subsets to reduce overfitting, while AdaBoost iteratively reweights misclassified instances to minimize bias (Cao Y, et al., 2013; Zhu LQ et al., 2017; Bayar B and Stamm MC, 2018).

To represent non-ensemble baselines, the authors incorporated KNN, SVM and SVR models. These models rely on local similarity (KNN) or kernel-based optimization (SVM/SVR), providing a useful contrast to tree-based learners, especially in cases with limited sample size (Drucker H et al., 1996; Vapnik V, 1998; Smola AJ and Schölkopf B, 2004; Marjanovic M et al., 2011; Nguyen B et al., 2016). A MLP was used to represent feedforward neural network-based learning (Rumelhart DE et al., 1986; Kuremoto T et al., 2014; Goodfellow I et al., 2016). Gaussian Process models (GPR, GPC) were added to test probabilistic learning approaches capable of expressing uncertainty; however, their high computational demands limited scalability in the regression task (Drineas P et al., 2005; Williams CK and Rasmussen CE, 2006; Murphy KP, 2012).

This diverse model set enables a comprehensive comparison across major supervised learning paradigms — ensemble, kernel-based, instance-based, neural and probabilistic—within a unified experimental framework.

3.2.3. Evaluation metrics

To comprehensively evaluate the performance of different machine learning classification models, the authors employ the following key metrics.

(i) Classification metrics. Accuracy represents the proportion of correctly classified samples. True Positive (*TP*) and True Negative (*TN*) denote correctly identified hydrate-bearing and non-hydrate-bearing data points, respectively, while False Positive (*FP*) and False Negative (*FN*) indicate misclassified samples. The calculation formula (Eq. 3) is presented as:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{3}$$

Precision measures the proportion of correctly predicted hydrate-bearing samples among all samples classified as hydrate-bearing. It is calculated as Eq. 4:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall represents the proportion of correctly identified hydrate-bearing samples out of all actual hydrate-bearing samples. The formula (Eq. 5) is presented as:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

The F1–score is the weighted harmonic mean of precision and recall, balancing the classifier's performance. It is computed as (Eq. 6):

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

Area Under the Curve (AUC) measures the model's classification performance across different thresholds. A higher AUC value (closer to 1) indicates a stronger classification ability.

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

where the true positive rate (*TPR*) is the proportion of actual hydrate-bearing samples that are correctly identified, and is therefore equivalent to *Recall* (Eq. 5); the false positive rate (*FP*R) is the proportion of actual hydrate-free samples that are incorrectly classified as hydrate-bearing.

(ii) Regression metrics. For regression tasks, the following evaluation metrics are used to assess model performance.

MSE measures the average squared difference between predicted and actual values. A smaller MSE indicates lower model error (Eq. 9).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{9}$$

Root Mean Squared Error (RMSE) is the square root of MSE, sharing the same unit as the target variable, making it easier to interpret the absolute error magnitude. A lower RMSE indicates better predictive accuracy (Eq. 10).

$$RMSE = \sqrt{\{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{10}$$

Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual values. Unlike MSE, it does not amplify larger errors disproportionately (Eq. 11):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \tag{11}$$

The R^2 score (Coefficient of Determination) reflects the model's ability to explain the variance in the data, with values interpreted as follows (Eq. 12):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - Y_i)^2} \tag{12}$$

$R^2 = 1$: The model perfectly fits the data.

$R^2 = 0$: The model performs at the level of mean-based prediction.

$R^2 < 0$: The model performs worse than mean-based prediction.

3.2.4. Hyperparameter optimization

Hyperparameter tuning is essential for balancing model complexity, error minimization, and generalization (Raschka S and Mirjalili V, 2019). This study employed grid search (GridSearchCV) for supervised learning models in both classification and regression tasks. Optimal parameters are summarized in Table S1.

(i) Classification Task optimization

For classification (hydrate occurrence identification), the objective was to maximize accuracy while avoiding overfitting.

Tree-based models (DT, RF, GBDT, XGBoost, LightGBM, CatBoost, AdaBoost, Bagging) were optimized for tree depth (max_depth), learning rate (learning_rate), and number of iterations (n_estimators). Regularization terms (min_samples_leaf, min_samples_split) were tuned to improve generalization.

Instance-based learning (KNN) required tuning of the number of neighbors (n_neighbors), balancing bias-variance trade-offs —lower values captured local structures, while higher values smoothed predictions.

Kernel-based models (SVM) used an RBF kernel, with regularization (C) adjusted for misclassification penalties and class_weight optimized for handling class imbalances.

Neural networks (MLP) were tuned for hidden layer architecture (hidden_layer_sizes), activation functions, and solvers to enhance convergence stability.

(ii) Regression Task optimization

For regression (hydrate saturation prediction), hyperparameter tuning focused on minimizing error and improving stability.

Decision tree and ensemble models (DT, RF, GBDT, XGBoost, CatBoost, LightGBM) were optimized for n_estimators, max_depth, and learning_rate, ensuring adequate complexity control through min_samples_leaf.

Gaussian Process Regression (GPR) was optimized by selecting an appropriate kernel and tuning length_scale to control smoothness. KNN regression required adjusting n_neighbors to balance sensitivity to local variations and overall prediction stability. Support Vector Regression (SVR) optimization focused on C for regularization and epsilon to define the tolerance margin, ensuring robustness against minor deviations. Multilayer Perceptron (MLP) regression was tuned for hidden_layer_sizes and solver (Adam) to enhance gradient update stability and improve convergence.

4. Results and interpretation

4.1. Performance of classification models

To evaluate the effectiveness of various machine learning classifiers in identifying gas hydrate-bearing formations, twelve models were tested under two labeling frameworks: Archie-based and velocity-derived. Table 1 summarizes their classification performance based on five-fold cross-validation and test set evaluation across six standard metrics.

4.1.1. Archie-based label results

Under the Archie-derived label configuration, ensemble methods achieved the highest classification performance. AdaBoost and GBDT both reached the highest test accuracy of 0.9444, followed closely by Bagging (0.9296) and Random Forest (0.9259). These models also achieved high F1–scores (⩾0.90 for AdaBoost and GBDT), suggesting a well-balanced classification between hydrate-bearing and hydrate-free intervals.

CatBoost, XGBoost, and LightGBM showed slightly lower test accuracy (0.91–0.92) but maintained strong AUC scores (>0.96), reflecting good model calibration. Decision Tree and KNN achieved moderate test accuracy (≈ 0.907), but with somewhat lower F1–scores (0.83), suggesting less robustness.

On the lower end, SVM, MLP, and GPC had lower test set accuracy (0.863–0.881) and comparatively weaker F1–scores (0.73–0.77), indicating difficulty in learning reliable hydrate classification boundaries from the given inputs.

Fig. 6 presents the receiver operating characteristic (ROC) curves of all models under this configuration, where ensemble methods show clearly steeper curves and larger areas under the curve, confirming their superior separability. Fig. 7 further illustrates the distribution of misclassifications through confusion matrices.

Table 1. Performance of classification models under Archie-based and velocity-derived label frameworks.

Model	Cross-validation accuracy (Mean±Std)		Test set accuracy		Precision (class 1)		Recall (class 1)		F1–score (class 1)		AUC	
	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label
Adaboost	0.9496±0.0109	0.9910±0.0104	0.9444	0.9888	0.92	0.99	0.89	0.95	0.90	0.98	0.9854	0.9918
Bagging	0.9377±0.0097	0.9899±0.0083	0.9296	0.9831	0.92	0.99	0.84	0.93	0.87	0.96	0.9844	0.9978
CatBoost	0.9199±0.0069	0.9921±0.0084	0.9185	0.9888	0.91	0.99	0.80	0.95	0.85	0.98	0.973	0.9997
Decision Tree	0.9303±0.0144	0.9697±0.0200	0.9074	0.9831	0.89	0.98	0.78	0.95	0.83	0.97	0.9126	0.9878
GBDT	0.9444±0.0096	0.9876±0.0120	0.9444	0.9888	0.91	0.99	0.90	0.95	0.90	0.98	0.9763	0.9983
GPC	0.8739±0.0195	0.9764±0.0090	0.8704	0.9775	0.83	0.99	0.70	0.91	0.76	0.95	0.9389	0.9908
KNN	0.8984±0.0072	0.9652±0.0115	0.9074	0.9775	0.90	0.99	0.77	0.91	0.83	0.95	0.9318	0.9849
LightGBM	0.9265±0.0122	0.9888±0.0118	0.9037	0.9831	0.88	0.99	0.77	0.93	0.82	0.96	0.9696	0.9941
MLP	0.8783±0.0132	0.7506±0.0028	0.863	0.6742	0.88	0.41	0.62	0.73	0.73	0.52	0.9375	0.7868
Random Forest	0.9325±0.0086	0.9888±0.0118	0.9259	0.9888	0.92	0.99	0.82	0.95	0.87	0.98	0.9801	0.9992
SVM	0.8776±0.0220	0.8876±0.0178	0.8815	0.9157	0.89	0.89	0.68	0.75	0.77	0.81	0.9327	0.9815
XGBoost	0.9288±0.0125	0.9876±0.0125	0.9111	0.9944	0.89	0.99	0.80	0.98	0.84	0.99	0.9671	0.9995

4.1.2. Velocity-derived label results

When trained using the velocity-derived label, most classifiers showed significantly higher performance, with nearly all ensemble models achieving test accuracy above 0.98. XGBoost attained the highest test set accuracy (0.9944), along with near-perfect precision (0.99) and an F1–score of 0.99, indicating extremely strong predictive capability under this label framework.

CatBoost, GBDT, Random Forest, and AdaBoost also recorded excellent results with F1–scores around 0.98, and AUC values exceeding 0.99. Simpler tree-based models like Decision Tree and Bagging similarly maintained high test accuracy (⩾0.9831), showing minimal drop-off in performance despite their reduced complexity.

KNN and GPC achieved solid generalization, with F1–scores of 0.95, while SVM showed modest test performance (F1 = 0.81). The MLP model underperformed substantially in this configuration, with a test accuracy of 0.6742 and an F1–score of only 0.52, indicating its limited applicability under this label scheme.

4.2. Performance of regression models

To estimate gas hydrate saturation (S_h) from well log and porewater chemistry data, twelve regression models were trained under both the Archie-based and velocity-derived label frameworks. Model performance was evaluated using cross-validation R^2 and three standard test-set metrics: MAE, RMSE and R^2. A full comparison is summarized in Table 2.

4.2.1. Archie-based label results

Under the Archie-based label configuration, Gaussian Process Regression (GPR) achieved the highest predictive accuracy, with a test R^2 of 0.9967 and RMSE as low as 0.0026, demonstrating its capacity to model fine-scale nonlinearities in the S_h distribution.

Several ensemble and instance-based models also performed strongly. GBDT, KNN, and Random Forest (RF) all produced test R^2 values above 0.93 and RMSE values under 0.012, suggesting they can robustly approximate the underlying physical relationships using limited training data. XGBoost and CatBoost followed closely, though slightly lagging in RMSE and generalization.

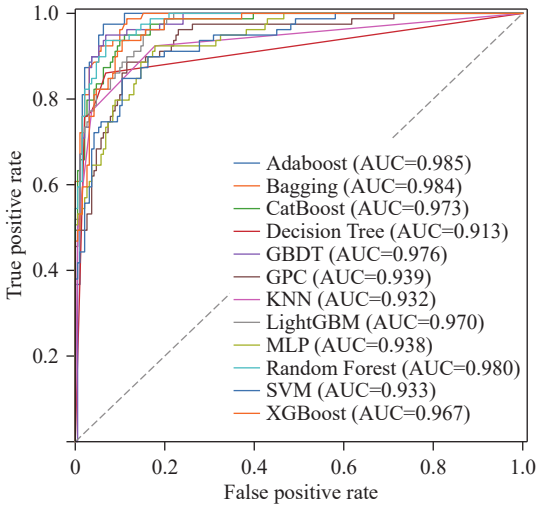

Fig. 6. Receiver operating characteristic (ROC) curves for twelve classification models under the Archie-based label configuration. The curves show the trade-off between true positive rate (TPR) and false positive rate (FPR) for binary classification of gas hydrate saturation (S_h). The diagonal dashed line represents the reference line (random guess). Area under the curve (AUC) values are indicated for each model; a higher AUC indicates better classification performance.

By contrast, models such as SVR and MLP performed markedly worse. The SVR model produced an R^2 of just 0.6345, while MLP's accuracy was even lower (R^2=0.477), indicating a limited ability to capture relevant patterns given the available feature space and sample size.

These findings are further visualized in Fig. 8, which shows the residual distribution for each model. Fig. 9, where predicted and actual saturation values are plotted, the most accurate models cluster tightly around the 1∶1 line, while lower-performing ones exhibit wider scatter and greater bias. Depth-wise performance is also illustrated in Fig. 10, where the predicted saturation profile shows strong alignment with

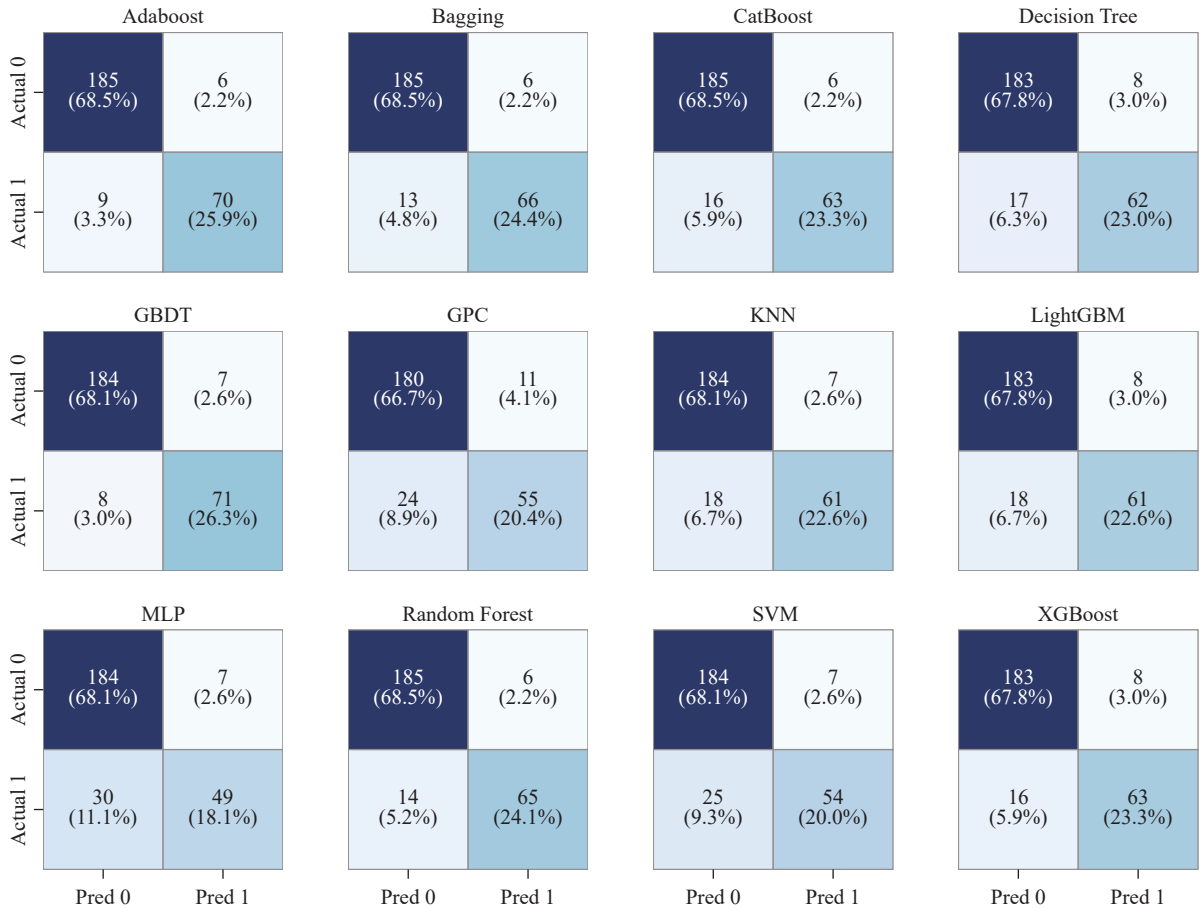

Fig. 7. Confusion matrices of twelve classification models under the Archie-based label configuration. Each subplot shows the number and proportion of correct and incorrect predictions for binary classification of gas hydrate saturation (S_h), with "0" representing non-hydrate intervals and "1" representing hydrate-bearing intervals.

hydrate-bearing zones.

4.2.2. Velocity-derived label results

Regression performance under the velocity-derived labels shows a wider spread, with top models still achieving high accuracy but others showing degradation, likely reflecting the reduced resolution and precision of this label set.

KNN and GPR again ranked among the most accurate models, with test R^2 exceeding 0.999 and RMSE values below 0.04. Tree-based ensemble methods such as GBDT, XGBoost, and CatBoost also maintained high performance, reaching R^2 values around 0.99. However, certain models, notably Bagging, RF, and LightGBM, exhibited higher RMSE despite high R^2, indicating increased sensitivity to label coarseness and outlier behavior.

The worst results were again observed in SVR and MLP. Under this label, MLP's test RMSE exceeded 0.99 with an R^2 below 0.5, and SVR showed large prediction errors with an RMSE of 1.23, confirming their limited applicability in this geophysical regression context.

5. Discussion

5.1. Analysis of model performance in the classification task

The classification results presented in Section 4.1 reveal a consistent pattern: Ensemble learning methods—particularly GBDT, Bagging, AdaBoost, and Random Forest—consistently achieve higher accuracy, F1–score, and AUC across both the Archie-based and velocity-derived label frameworks (Breiman L, 2001). This performance advantage is attributable to the structural characteristics of these models and their compatibility with the complexity of gas hydrate identification.

Gas hydrate-bearing layers exhibit complex, non-linear signatures in geophysical logs due to their interactions with acoustic and resistivity properties (Collett TS and Ladd J, 2000; Shankar U and Riedel M, 2014). Ensemble methods are structurally well-suited to model such complexities. In particular, GBDT and its variants combine weak learners sequentially through gradient descent on residual errors, allowing them to refine predictions with high granularity (Chen TQ and Guestrin C, 2016; Ke GL et al., 2017). Their ability to perform feature-wise splits and capture subtle interactions makes them effective in distinguishing hydrate-bearing from hydrate-free interval, especially when features such as V_P, V_S, or resistivity exhibit threshold-like changes near the base of stability zones.

Bagging-based methods, including Random Forest and Bagging itself, leverage variance reduction by aggregating

Table 2. Performance of regression model under Archie-based and velocity-derived label frameworks.

Model	Cross-validation R^2 (Mean±Std)		Test Set MAE		Test Set RMSE		Test Set R^2	
	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label	Archie-based label	Velocity-derived label
Adaboost	0.9201 ± 0.0262	0.9922 ± 0.0033	0.0067	0.0122	0.0098	0.1621	0.9520	0.9861
Bagging	0.9265 ± 0.0270	0.9476 ± 0.0350	0.0069	0.1322	0.0102	0.3114	0.9479	0.9486
CatBoost	0.8873 ± 0.0268	0.9748 ± 0.0173	0.0091	0.0332	0.0138	0.1255	0.9046	0.9917
Decision Tree	0.8858 ± 0.0231	0.9678 ± 0.0386	0.0081	0.0081	0.0124	0.1081	0.9227	0.9938
GBDT	0.9243 ± 0.0221	0.9664 ± 0.0260	0.0065	0.0280	0.0092	0.1315	0.9573	0.9908
GPR	0.9936 ± 0.0024	0.9714 ± 0.0370	0.0016	0.0038	0.0026	0.0369	0.9967	0.9993
KNN	0.9718 ± 0.0108	0.9808 ± 0.0168	0.0039	0.0037	0.0062	0.0324	0.9804	0.9994
LightGBM	0.8755 ± 0.0299	0.9613 ± 0.0254	0.0104	0.0918	0.0154	0.2597	0.8800	0.9643
MLP	0.4531 ± 0.0404	0.6304 ± 0.1145	0.0233	0.5719	0.0322	0.9945	0.4770	0.4760
Random Forest	0.9263 ± 0.0272	0.9478 ± 0.0347	0.0068	0.1311	0.0102	0.3088	0.9480	0.9495
SVR	0.6222 ± 0.0456	0.3956 ± 0.1579	0.0184	0.4692	0.0269	1.2313	0.6345	0.1968
XGBoost	0.8881 ± 0.0388	0.9643 ± 0.0312	0.0084	0.0326	0.0118	0.1447	0.9293	0.9889

multiple decorrelated trees trained on bootstrapped samples (Breiman L, 1996). This strategy yields robust generalization, particularly in settings like this where sample size is limited (1350 samples for Archie-based, 890 samples for velocity-based), and label noise or variability is inevitable. The high AUCs and consistent precision/recall balance observed for these models reflect their capacity to handle both feature diversity and label uncertainty.

In contrast, SVM and MLP underperformed across both labeling configurations, suggesting structural mismatches with the problem space. SVMs rely on maximizing margin between classes in a high-dimensional feature space, but are sensitive to kernel choice and hyperparameter settings. Given the small dataset size and heterogeneity in signal characteristics, they may struggle to find stable decision boundaries. Similarly, MLP, while theoretically expressive, are prone to overfitting on limited tabular data unless carefully regularized. Their lower F1–scores and higher error variance in this study point to insufficient learning of class-specific feature interactions (Goodfellow I et al., 2016).

5.2. Analysis of model performance in the regression task

In the regression task targeting hydrate saturation (S_h) estimation, clear performance stratification is again observed across the different machine learning models. The results align closely with the theoretical underpinnings and computational properties of each model (Raschka S and Mirjalili V, 2019).

Gaussian Process Regression (GPR) achieves the highest overall performance under both label frameworks, with an R^2 of 0.9967 in the Archie-based configuration and 0.9993 in the velocity-derived configuration. GPR's strength lies in its ability to model uncertainty and capture complex nonlinear trends, even in relatively small datasets, by constructing a distribution over possible functions (Williams CK and Rasmussen CE, 2006). However, its scalability remains a limitation, especially for larger datasets or real-time applications, given its cubic computational complexity.

Ensemble methods, particularly GBDT and RF, consistently perform well, with R^2 values exceeding 0.93 under the Archie label and above 0.99 under the velocity-derived label. GBDT and RF leverage ensemble averaging and residual fitting to manage variance and bias trade-offs, making them highly suited to the moderate sample sizes and nonlinear feature–target relationships in this study. KNN, while conceptually simple, benefits from strong local adaptivity when the sample density is sufficient and noise is moderate, explaining its strong showing here.

On the other hand, MLP and SVR underperform, particularly under the Archie-based label, with R^2 values of 0.477 and 0.635, respectively. This is likely due to their need for larger training data to generalize effectively. The MLP, despite its theoretical capacity for universal function approximation, may be hindered by overfitting or local minima during optimization given the limited dataset size. SVR, while robust to noise, depends heavily on kernel and parameter tuning, and may not be well-matched to the multi-scale variability of hydrate saturation features (Goodfellow I et al., 2016).

Interestingly, XGBoost and LightGBM, despite being advanced gradient boosting frameworks, did not outperform classic GBDT or RF in regression accuracy. This could be attributed to the relatively modest sample size and the models' reliance on fine-grained parameter tuning, which may require more extensive optimization than performed here.

While prediction accuracy is critical, practical deployment in field scenarios also requires consideration of computational efficiency. As summarized in Table S2, GPR stands out for its high memory and time requirements, approximately 6 times slower and 10 times more memory-intensive than GBDT. While GPR is viable for exploratory analysis or small datasets, it may be impractical in large-scale applications (Williams CK and Rasmussen CE, 2006).

In contrast, models like GBDT and Random Forest maintain strong predictive power ($R^2 > 0.93$) with far lower resource demands, making them more scalable. KNN also offers high accuracy with modest runtime, though it may be sensitive to sample distribution.

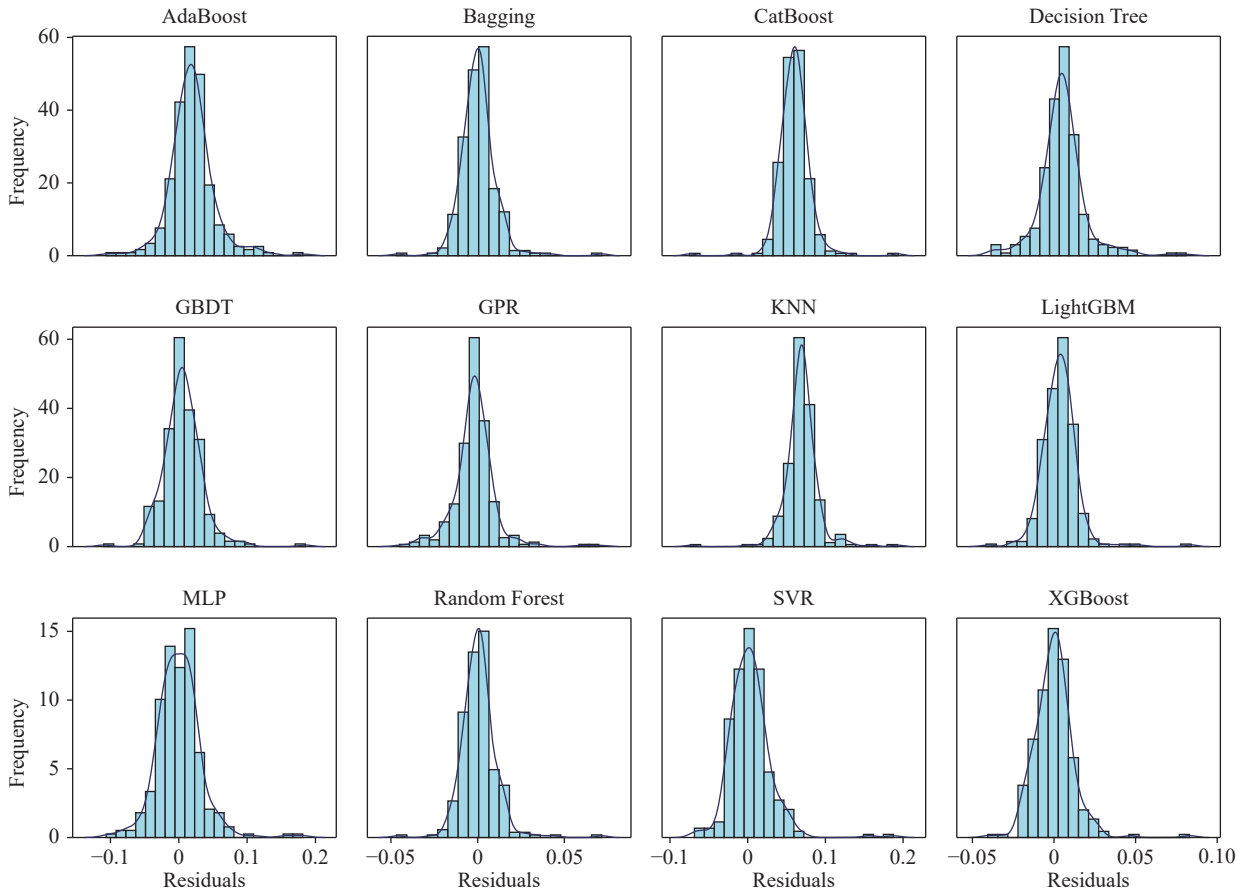

Fig. 8. Residual distributions of 12 regression models under the Archie-based label configuration. Each subplot shows a histogram of residuals (prediction − Archie-derived gas hydrate saturation, S_h) overlaid with a kernel density estimate (KDE) curve. The distributions illustrate model bias and error spread relative to the Archie-derived reference.

These comparisons underline the importance of selecting models not only for accuracy, but also for computational feasibility in field applications.

5.3. Generalization and feature interpretability

5.3.1. Group-based validation and overlooked data leakage risks

A frequently overlooked challenge in geoscientific machine learning is data leakage introduced by random-sample cross-validation (Michelucci U, 2024). When consecutive depth measurements are randomly split across training and test sets, models may exploit autocorrelated features rather than learning generalizable geological patterns (Karpatne A et al., 2018; Zuo RG, 2017). This can lead to overly optimistic estimates of predictive accuracy, especially in stratified subsurface datasets.

To mitigate this risk, the authors adopted a group-based validation scheme that divides the dataset into contiguous 3-meter intervals, ensuring group-wise exclusivity between training and test sets. The performance of classification models under this setup is reported in Table S3.

When compared to the standard random-split results in Table 1, most models experienced a reduction in test accuracy of approximately 18%–20% points, and a decline in AUC by about 13–16 points. For instance, GBDT dropped from 0.9444 to 0.7482 in test accuracy and from 0.9763 to 0.8503 in AUC; Random Forest declined from 0.9259 to 0.7445 in test accuracy, and from 0.9801 to 0.8383 in AUC. These shifts confirm that part of the predictive signal in random splits likely stemmed from intra-sample correlation (Zhu LQ et al., 2021).

Nevertheless, even under the stricter group-based validation, ensemble models such as XGBoost, AdaBoost, GBDT, and Random Forest maintained F1–scores in the range of 0.69–0.72 and AUCs exceeding 0.82. This suggests that while predictive certainty diminishes in more realistic deployment settings, the models retain a core capacity for hydrate layer identification.

In practical geophysical workflows, such declines may be partially mitigated by incorporating additional geological inputs such as temperature gradients and lithological indicators, or through more aggressive hyperparameter tuning. Importantly, group-based validation offers a more realistic lower-bound on model robustness, aligning more closely with expected performance in unseen stratigraphic contexts or nearby wells.

5.3.2. Cross-label generalization and feature variation robustness

As shown in Section 4, both classification and regression

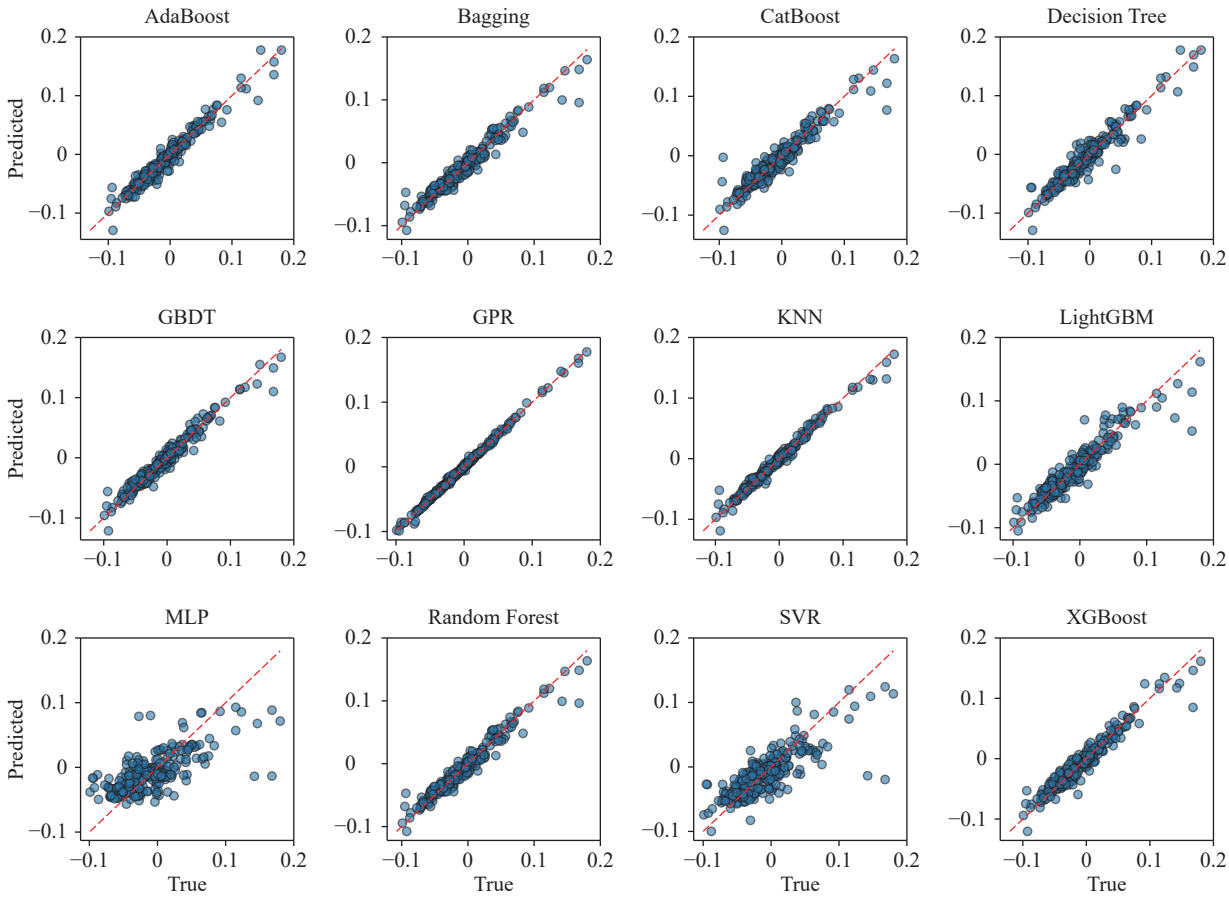

Fig. 9. Scatter plots comparing predicted and Archie-derived gas hydrate saturation (S_h) values for 12 regression models under the Archie-based label configuration. Each subplot corresponds to a different model. The red dashed line represents the reference line ($y = x$); points closer to this line indicate better model performance.

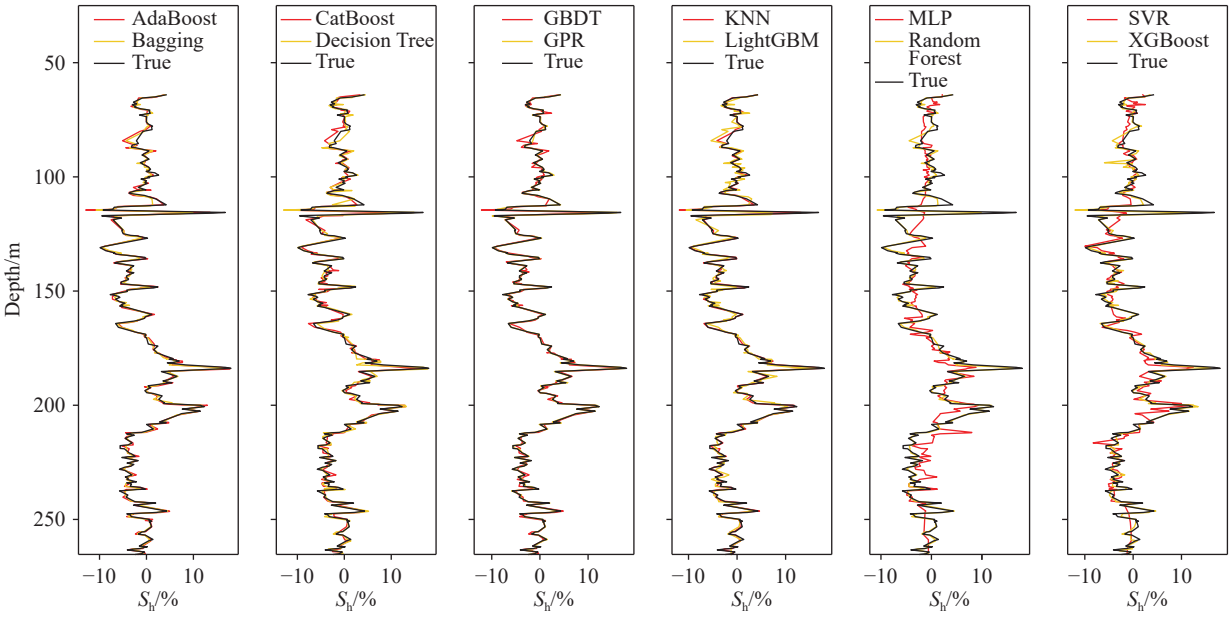

Fig. 10. Predicted and Archie-derived gas hydrate saturation (S_h) profiles in Well NGHP-01-19B under the Archie-based label configuration. The Archie-derived values are shown in black, and the model predictions are plotted in red and gold.

models exhibit a high degree of consistency in performance rankings across the two labeling frameworks, Archie-based and velocity-derived, despite differences not only in target definitions but also in input feature sets.

For classification, ensemble models such as XGBoost, GBDT, AdaBoost, and Random Forest consistently

outperformed others under both labeling schemes, achieving the highest test accuracies and F1–scores. In contrast, models like SVM and MLP showed consistently weaker performance, particularly in recall and F1–score.

In the regression task, GPR maintained the top accuracy under both label types, followed closely by GBDT, KNN, and Random Forest. Again, MLP and SVR underperformed across the board.

Importantly, the shift between labeling schemes was accompanied by a natural change in input features: The Archie-based configuration used acoustic log data (V_P, V_S, and TDEP), while the velocity-derived setup employed resistivity and pore-water geochemical features (chloride and salinity). This was not an artificial setup, but rather a consequence of preserving physical separation between features and target variables (Section 3.2.1).

The same models performed best across both configurations, despite these fundamental differences. This indicates that they are not merely exploiting dataset-specific statistical artifacts. Instead, they demonstrate the ability to extract generalizable signals related to hydrate-bearing formation characteristics across distinct physical data domains (Zuo RG, 2017; Karpatne A et al., 2018).

This cross-label, cross-feature robustness underscores the adaptability of these models for hydrate exploration scenarios, where input data types may vary due to survey design, formation heterogeneity, or data availability.

5.3.3. Implicit significance of TDEP as a contextual proxy

In the Archie-based label configuration, the input feature TDEP, which represents true vertical depth, may initially appear to carry little direct geophysical relevance to hydrate occurrence, as it lacks the physical interpretability of parameters like resistivity or porosity. However, further ablation experiments reveal that removing TDEP results in a significant degradation of model performance across the board.

Specifically, for top-performing classifiers such as GBDT, Bagging, XGBoost, and CatBoost, test accuracy dropped from 0.9444 to 0.7259, 0.9296 to 0.7370, 0.9111 to 0.7370, and 0.9185 to 0.7556 respectively, representing relative declines of approximately 20%–22%. This pattern is consistent across other ensemble and baseline models as well, with most experiencing a decline in test accuracy in the range of 17%–23% (Table S4).

These results suggest that while TDEP itself does not directly encode petrophysical or geochemical information, it likely serves as a proxy for broader geological conditions, especially ambient pressure and thermal gradients governing the gas hydrate stability zone (GHSZ). In marine settings, hydrate-bearing layers often exhibit stratigraphic confinement within specific depth intervals (Max MD and Johnson AH, 2019; Wang JJ and Lau HC, 2020). Therefore, depth can implicitly signal the likelihood of hydrate occurrence, especially when combined with other features.

This finding underscores the importance of not dismissing seemingly generic features, as they may embody latent domain knowledge through their spatial correlation with geological processes. In future applications, the role of such proxy variables warrants careful consideration, especially when designing feature sets in data-sparse environments.

6. Conclusions

This study presents a comprehensive assessment of supervised machine learning algorithms for gas hydrate identification and saturation estimation in marine reservoirs, using well log and pore-water geochemical data from Site NGHP-01-19B. Based on experimental findings and validation analysis, the following conclusions are drawn.

(i) Ensemble learning models, especially Gradient Boosting Decision Trees (GBDT), Random Forest, AdaBoost, and XGBoost, demonstrated superior classification accuracy under both resistivity-based (Archie's law) and velocity-based label (three-phase weighted equation) frameworks. These models achieved test accuracies exceeding 0.94 under standard validation and retained robust F1–scores and AUCs under group-based validation, confirming their reliability in handling heterogeneous and nonlinear hydrate-related geophysical signals.

(ii) For regression tasks, Gaussian Process Regression (GPR) achieved the highest accuracy in predicting hydrate saturation ($R^2 > 0.99$), though its scalability is limited. Tree-based ensemble models such as GBDT and Random Forest offered a practical balance between predictive accuracy and computational efficiency, making them suitable for large-scale or real-time hydrate evaluation scenarios.

(iii) The consistency of model performance across two physically independent labeling frameworks, paired with distinct input feature sets, demonstrates strong cross-label generalizability. Additionally, group-based validation highlighted the risk of overfitting due to depth-wise autocorrelation in random splits. Features such as depth below seafloor (TDEP), although not directly geophysical, contributed significantly to model accuracy by acting as contextual proxies for stratigraphic and thermodynamic conditions.

These findings affirm the viability of machine learning as a robust and interpretable toolset for offshore gas hydrate reservoir characterization. Future extensions may include multi-well integration, seismic attribute fusion, or incorporation of semi-supervised and deep learning models to enhance performance under data-limited conditions.

CRediT authorship contribution statement

Yi-fan Wu led the study design, conducted all experiments and data analysis, and wrote the manuscript. Zheng Su provided project funding, supervision, and academic guidance. Takeshi Tsuji contributed as an international advisor. Dai-dai Wu, Chuang-ji Feng, Guang-rong Jin, and Chao Yang participated in project coordination and discussion. Neng-you Wu contributed to scientific discussion and general guidance. All authors reviewed and approved the final manuscript.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgment

This research was supported by the China Scholarship Council under the State Scholarship Fund (202506340082), the Key Project of Guangdong Provincial Key R & D Program (2023B1111050014), the Youth Promotion Project of the Natural Science Foundation of Guangdong Province (2023A1515030280), the Guangdong Basic and Applied Basic Research Foundation (2023A1515010926), and the Guangzhou Science and Technology Plan Project (2024A04J9876). Additional support was funded by China National Petroleum Corporation (CNPC, 2024DQ02-0107).

The author gratefully acknowledges the data support provided by the National Gas Hydrate Program Expedition 01 (NGHP-01), jointly organized by the Directorate General of Hydrocarbons (DGH), Ministry of Petroleum and Natural Gas (India), and the Consortium for Scientific Methane Hydrate Investigations (CSMHI). The logging data used in this study were acquired using the drilling platform JOIDES Resolution, with technical support from IODP, USDOE, TAMU and FUGRO.

The authors sincerely thank the three reviewers for their constructive comments and suggestions, which significantly improved the quality and clarity of this manuscript.

Supplementary dataset

The data in this paper can be accessed through http://www-odp.tamu.edu/. Supplementary data (Tables S1–S4) to this article can be found online at doi: 10.31035/cg2025078..

References

Ahmad AB, Tsuji T. 2021. Machine learning for automatic slump identification from 3D seismic data at convergent plate margins. Marine and Petroleum Geology, 133, 105290. doi: 10.1016/j.marpetgeo.2021.105290.

Archie GE. 1942. The electrical resistivity log as an aid in determining some reservoir characteristics. Transactions of the AIME, 146(1), 54–62. doi: 10.2118/942054-G.

Bastia R. 2006. An overview of Indian sedimentary basins with special focus on emerging east coast deepwater frontiers. The Leading Edge, 25(7), 818–829. doi: 10.1190/1.2221359.

Bastia R, Radhakrishna M, Das S, Kale AS, Catuneanu O. 2010. Delineation of the 85 E ridge and its structure in the Mahanadi Offshore Basin, Eastern Continental Margin of India (ECMI), from seismic reflection imaging. Marine and Petroleum Geology, 27(9), 1841–1848. doi: 10.1016/j.marpetgeo.2010.08.003.

Bayar B, Stamm MC. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security, 13(11), 2691–2706. doi: 10.1109/TIFS.2018.2825953.

Bharali B. 1991. A brief review of Mahanadi delta and the deltaic sediments in Mahanadi basin. Memoir Geological Society of India, 22, 31–49.

Breiman L. 1996. Bagging predictors. Machine Learning, 24(2), 123–140. doi: 10.1007/BF00058655.

Breiman L. 2001. Random forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324.

Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, Vanderplas J, Joly A, Holt B, Varoquaux G. 2013. API design for machine learning software: Experiences from the scikit-learn project. arXiv, 1309.0238. doi: 10.48550/arXiv.1309.0238.

Cao Y, Miao QG, Liu JC, Gao L. 2013. Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 39(6), 745–758. doi: 10.1016/S1874-1029(13)60052-X.

Chen TQ, Guestrin C. 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. doi: 10.1145/2939672.2939785.

Chong L, Collett TS, Creason CG, Seol Y, Myshakin EM. 2024. Machine-learning application to assess occurrence and saturations of methane hydrate in marine deposits offshore India. Interpretation, 12(1), T63–T75. doi: 10.1190/INT-2023-0056.1.

Chong L, Singh H, Creason CG, Seol Y, Myshakin EM. 2022. Application of machine learning to characterize gas hydrate reservoirs in Mackenzie Delta (Canada) and on the Alaska north slope (USA). Computational Geosciences, 26(5), 1151–1165. doi: 10.1007/s10596-022-10151-9.

Collett TS, Riedel M, Cochran JR, Boswell R, Kumar P, Sathe AV. 2008. Indian continental margin gas hydrate prospects: Results of the Indian National Gas Hydrate Program (NGHP) Expedition 01. doi: 10.7916/D8-NF9W-CZ91.

Collett TS. 2002. Energy resource potential of natural gas hydrates. AAPG Bulletin, 86(11), 1971–1992. doi: 10.1306/61EEDDD2-173E-11D7-8645000102C1865D.

Collett TS, Boswell R, Cochran JR, Kumar P, Lall M, Mazumdar A, Ramana MV, Ramprasad T, Riedel M, Sain K, Sathe AV, Vishwanath K. 2014. Geologic implications of gas hydrates in the offshore of India: Results of the National Gas Hydrate Program Expedition 01. Marine and Petroleum Geology, 58, 3–28. doi: 10.1016/j.marpetgeo.2014.07.021.

Collett TS, Ladd J. 2000. Detection of gas hydrate with downhole logs and assessment of gas hydrate concentrations (saturations) and gas volumes on the Blake Ridge with electrically resistivity log data. In: Proceedings of the Ocean Drilling Program, Scientific Results, 164. College Station, TX, Texas A & M University. https://www.academia.edu/download/70992860/227b90d630a066a2f498f64fac64e6a1c3f7.pdf.

Collett TS, Riedel M, Boswell R, Presley J, Kumar P, Sathe A, Sethi A, Lall MV, NGHP Expedition scientists. 2015. Indian National Gas Hydrate Program Expedition 01 report. Reston, US Geological Survey. https://pubs.usgs.gov/publication/sir20125054.

Dangwal V, Sengupta S, Desai AG. 2008. Speculated petroleum systems in deep offshore Mahanadi Basin in Bay of Bengal, India. International Conference & Exposition on Petroleum Geophysics, Hyderabad, India, 1–7. https://spgindia.org/2008/001.pdf.

Dickens GR. 2003. Rethinking the global carbon cycle with a large, dynamic and microbially mediated gas hydrate capacitor. Earth and Planetary Science Letters, 213(3–4), 169–183. doi: 10.1016/S0012-821X(03)00325-X.

Drineas P, Mahoney MW, Cristianini N. 2005. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. Journal of Machine Learning Research, 6(12), 2153–2175. http://www.jmlr.org/papers/volume6/drineas05a/drineas05a.pdf.

Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. 1996. Support vector regression machines. Advances in Neural Information Processing Systems, 9, 155–161. https://proceedings.neurips.cc/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html.

Fofonoff NP. 1985. Physical properties of seawater: A new salinity scale and equation of state for seawater. Journal of Geophysical Research:

Oceans, 90(C2), 3332–3342. doi: 10.1029/JC090iC02p03332.

Fuloria RC, Pandey RN, Bharali BR, Mishra JK. 1992. Stratigraphy, structure and tectonics of Mahanadi offshore basin. Visesa Prakasana-Bharatiya Bhuvaijñanika Sarveksana, 29, 255–265. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=6375318.

Goodfellow I, Bengio Y, Courville A. 2016. Deep feedforward networks. Deep Learning, 1, 161–217.

Hancock JT, Khoshgoftaar TM. 2020. CatBoost for big data: An interdisciplinary review. Journal of Big Data, 7(1), 94. doi: 10.1186/s40537-020-00369-8.

Harris JR, Grunsky EC 2015. Predictive lithological mapping of Canada's north using Random Forest classification applied to geophysical and geochemical data. Computers & Geosciences, 80, 9–25. doi: 10.1016/j.cageo.2015.03.013.

Jagannathan CR, Ratnam C, Baishva NC, Gupta UD. 1984. Geology of the offshore Mahanadi basin. Petroleum Asia Journal, 6(4), 101–104. https://www.osti.gov/etdeweb/biblio/6776553.

Jana S, Ojha M, Sain K, Srivastava S. 2017. An approach to estimate gas hydrate saturation from 3-D heterogeneous resistivity model: A study from Krishna-Godavari Basin, Eastern Indian offshore. Marine and Petroleum Geology, 79, 99–107. doi: 10.1016/j.marpetgeo.2016.11.006.

Karpatne A, Ebert-Uphoff I, Ravela S, Babaie HA, Kumar V. 2018. Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544–1554. doi: 10.1109/TKDE.2018.2861006.

Ke GL, Meng Q, Finley T, Wang TF, Chen W, Ma WD, Ye QW, Liu TY. 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Kuremoto T, Kimura S, Kobayashi K, Obayashi M. 2014. Time series forecasting using a deep belief network with restricted Boltzmann machines. Neurocomputing, 137, 47–56. doi: 10.1016/j.neucom.2013.03.047.

Ma XL, Ding CD, Luan SL, Wang Y, Wang YP. 2017. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Transactions on Intelligent Transportation Systems, 18(9), 2303–2310. doi: 10.1109/TITS.2016.2635719.

Malinverno A, Kastner M, Torres ME, Wortmann UG. 2008. Gas hydrate occurrence from pore water chlorinity and downhole logs in a transect across the northern Cascadia margin (Integrated Ocean Drilling Program Expedition 311). Journal of Geophysical Research: Solid Earth, 113(B8), B08103. doi: 10.1029/2008JB005702.

Marjanovic M, Kovacevic M, Bajat B, Vozenilek V. 2011. Landslide susceptibility assessment using SVM machine learning algorithm. Engineering Geology, 123(3), 225–234. doi: 10.1016/j.enggeo.2011.09.006.

Max MD, Johnson AH. 2019. Economic characteristics of deepwater natural gas hydrate. In: Max MD, Johnson AH (eds). Exploration and Production of Oceanic Natural Gas Hydrate, 57–94. Cham, Springer International Publishing. doi: 10.1007/978-3-030-00401-9_2.

Michelucci U. 2024. Fundamental Mathematical Concepts for Machine Learning in Science. Cham, Springer International Publishing. doi: 10.1007/978-3-031-56431-4.

Milkov AV. 2004. Global estimates of hydrate-bound gas in marine sediments: How much is really out there? Earth-Science Reviews, 66(3–4), 183–197. doi: 10.1016/j.earscirev.2003.11.002.

Milkov AV, Sassen R. 2003. Preliminary assessment of resources and economic potential of individual gas hydrate accumulations in the Gulf of Mexico continental slope. Marine and Petroleum Geology, 20(2), 111–128. doi: 10.1016/S0264-8172(03)00024-2.

Mohapatra P. 2006. Sequence stratigraphic approach for identification of hydrocarbon plays in Mahanadi offshore basin. 6th International Conference & Exposition on Petroleum Geophysics, 672–677. https://www.academia.edu/download/76798365/231.pdf.

Murphy KP. 2012. Machine Learning: A Probabilistic Perspective. Cambridge, MIT Press. https://books.google.com/books?hl=en&lr=&id=RC43AgAAQBAJ.

Nguyen B, Morell C, De Baets BD. 2016. Large-scale distance metric learning for k-nearest neighbors regression. Neurocomputing, 214, 805–814. doi: 10.1016/j.neucom.2016.07.005.

Nguyen H, Bui XN, Bui HB, Cuong DT. 2019. Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: A case study. Acta Geophysica, 67(2), 477–490. doi: 10.1007/s11600-019-00268-4.

Pearson CF, Halleck PM, McGuire PL, Hermes R, Mathews M. 1983. Natural gas hydrate deposits: A review of in situ properties. The Journal of Physical Chemistry, 87(21), 4180–4185. doi: 10.1021/j100244a041.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825–2830.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. 2018. CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Raschka S, Mirjalili V. 2019. Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2. Birmingham, Packt Publishing Ltd. https://books.google.com/books?hl=en&id=sKXIDwAAQBAJ.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. Nature, 323(6088), 533–536. doi: 10.1038/323533a0.

Sastri VV, Venkatachala BS, Narayanan V. 1981. The evolution of the east coast of India. Palaeogeography, Palaeoclimatology, Palaeoecology, 36(1–2), 23–54. doi: 10.1016/0031-0182(81)90047-X.

Shankar U, Riedel M. 2014. Assessment of gas hydrate saturation in marine sediments from resistivity and compressional-wave velocity log measurements in the Mahanadi Basin, India. Marine and Petroleum Geology, 58, 265–277. doi: 10.1016/j.marpetgeo.2013.10.007.

Singh H, Seol Y, Myshakin EM. 2021. Prediction of gas hydrate saturation using machine learning and optimal set of well-logs. Computational Geosciences, 25(1), 267–283. doi: 10.1007/s10596-020-10004-3.

Sloan ED. 1991. Natural gas hydrates. Journal of Petroleum Technology, 43(12), 1414–1417. doi: 10.2118/23562-PA.

Smola AJ, Schölkopf B. 2004. A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222. doi: 10.1023/B:STCO.0000035301.49549.88.

Spangenberg E. 2001. Modeling of the influence of gas hydrate content on the electrical properties of porous sediments. Journal of Geophysical Research: Solid Earth, 106(B4), 6535–6548. doi: 10.1029/2000JB900434.

Su Z, Cao YC, Wu NY, Chen DF, Yang SX, Wang HB. 2012. Numerical investigation on methane hydrate accumulation in Shenhu area, northern continental slope of South China Sea. Marine and Petroleum Geology, 38(1), 158–165. doi: 10.1016/j.marpetgeo.2012.06.005.

Su Z, Huang L, Wu NY, Yang SX. 2013. Effect of thermal stimulation on gas production from hydrate deposits in Shenhu area of the South China Sea. Science China Earth Sciences, 56(4), 601–610. doi: 10.1007/s11430-013-4587-4.

Sultan N, Cochonat P, Foucher JP, Mienert J. 2004. Effect of gas

hydrates melting on seafloor slope instability. Marine Geology, 213(1–4), 379–401. doi: 10.1016/j.margeo.2004.10.015.

Tian DM, Yang SX, Gong YH, Geng MH, Li YH, Hu G. 2023. A comparative study of machine learning methods for gas hydrate identification. Geoenergy Science and Engineering, 223, 211564. doi: 10.1016/j.geoen.2023.211564.

Vapnik V. 1998. The Support Vector Method of Function Estimation. In: Suykens JAK, Vandewalle J. Nonlinear Modeling: Advanced Black-Box Techniques. Boston, Springer, 55–85. doi: 10.1007/978-1-4615-5703-6_3.

Wang JJ, Lau HC. 2020. Thickness of gas hydrate stability zone in permafrost and marine gas hydrate deposits: Analysis and implications. Fuel, 282, 118784. doi: 10.1016/j.fuel.2020.118784.

Wang XJ, Wu SG, Lee M, Guo YQ, Yang SX, Liang JQ. 2011. Gas hydrate saturation from acoustic impedance and resistivity logs in the Shenhu area, South China Sea. Marine and Petroleum Geology, 28(9), 1625–1633. doi: 10.1016/j.marpetgeo.2011.07.002.

Williams CK, Rasmussen CE. 2006. Gaussian Processes for Machine Learning. Cambridge, MIT Press. https://www.newton.ac.uk/files/seminar/20070809140015001-150844.pdf.

Wood AB. 1941. A Textbook of Sound. New York, MacMillan.

Wyllie MRJ, Gregory AR, Gardner GHF. 1958. An experimental investigation of factors affecting elastic wave velocities in porous media. Geophysics, 23(3), 459–493. doi: 10.1190/1.1438493.

Xu HF, Jiao ZR, Zhang ZR, Huffman M, Wang QS. 2021. Prediction of methane hydrate formation conditions in salt water using machine learning algorithms. Computers & Chemical Engineering, 151, 107358. doi: 10.1016/j.compchemeng.2021.107358.

Yuan HM, Wang Y, Wang XC. 2021. Seismic methods for exploration and exploitation of gas hydrate. Journal of Earth Science, 32(4), 839–849. doi: 10.1007/s12583-021-1502-3.

Yuan T, Hyndman RD, Spence GD, Desmons B. 1996. Seismic velocity increase and deep-sea gas hydrate concentration above a bottom-simulating reflector on the northern Cascadia continental slope. Journal of Geophysical Research: Solid Earth, 101(B6), 13655–13671. doi: 10.1029/96JB00102.

Zhao XB, Chen XJ, Huang Q, Lan ZJ, Wang XG, Yao GQ. 2022. Logging-data-driven permeability prediction in low-permeable sandstones based on machine learning with pattern visualization: A case study in Wenchang A Sag, Pearl River Mouth Basin. Journal of Petroleum Science and Engineering, 214, 110517. doi: 10.1016/j.petrol.2022.110517.

Zhu LQ, Zhang C, Wei Y, Zhang CM. 2017. Permeability prediction of the tight sandstone reservoirs using hybrid intelligent algorithm and nuclear magnetic resonance logging data. Arabian Journal for Science and Engineering, 42(4), 1643–1654. doi: 10.1007/s13369-016-2365-2.

Zhu LQ, Zhou XQ, Zhang CM. 2021. Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. Artificial Intelligence in Geosciences, 2, 76–81. doi: 10.1016/j.aiig.2021.12.001.

Zhu LQ, Wei JG, Wu SG, Zhou XQ, Sun J. 2022. Application of unlabelled big data and deep semi-supervised learning to significantly improve the logging interpretation accuracy for deep-sea gas hydrate-bearing sediment reservoirs. Energy Reports, 8, 2947–2963. doi: 10.1016/j.egyr.2022.01.139.

Zhu LQ, Zhou XQ, Sun J, Liu YR, Wang JC, Wu SG. 2023. Reservoir classification and log prediction of gas hydrate occurrence in the Qiongdongnan Basin, South China Sea. Frontiers in Marine Science, 10, 1055843. doi: 10.3389/fmars.2023.1055843.

Zuo RG. 2017. Machine learning of mineralization-related geochemical anomalies: A review of potential methods. Natural Resources Research, 26(4), 457–464. doi: 10.1007/s11053-017-9345-4.