

Article

Prediction of Water Saturation Using Physics-Guided Machine Learning in Deep Silurian Shale Gas Reservoirs

Gaofeng Zou ^{1,2,3,4}, Liang Xue ^{1,2,*}, Haiyang Chen ^{1,2}, Ruyue Wang ^{4,5}, Yubin Dong ^{1,2}, Di Tian ^{1,2} and Minghao Wang ^{1,2}

¹ State Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing 102249, China

² Department of Oil-Gas Field Development Engineering, College of Petroleum Engineering, China University of Petroleum, Beijing 102249, China

³ Exploration and Development Research Institute of Jiangnan Oilfield Branch, Sinopec, Wuhan 430223, China

⁴ State Key Laboratory of Shale Oil and Gas Accumulation Mechanisms and Efficient Development, Beijing 102206, China

⁵ Research Institute of Petroleum Exploration and Development, Sinopec, Beijing 102206, China

* Correspondence: xueliang@cup.edu.cn

Abstract

Accurate water saturation estimation in deep shale reservoirs is complicated by clay-related additional conductivity and coupled pore, organic-matter, and structural effects. This study develops a feature-level physics-guided machine-learning framework, termed PhysML-Hybrid. Five mechanism-derived descriptor groups representing clay–water interfacial behavior, low-resistivity correction, pore connectivity, organic-pore development, and structural stress were integrated with conventional reservoir variables in a validation-weighted ensemble of random forest, XGBoost, and Bayesian neural network models. The framework was evaluated using 153 depth-matched samples from five wells in the Dingshan area of the Sichuan Basin. The data were divided into 107 training, 16 validation, and 30 independent test samples, and target-stratified five-fold cross-validation was conducted exclusively within the training set. Mean cross-validation R^2 , MAE, and RMSE were 0.907 ± 0.009 , $1.69\% \pm 0.10\%$, and $2.25\% \pm 0.14\%$, respectively. On the independent test set, the corresponding values were 0.902, 1.77%, and 2.34%. PhysML-Hybrid outperformed Archie, SVM, ML-only, and Phy-XGB. SHAP and statistical analyses identified clay content, the curvature–clay interaction, TOC, pore connectivity, and structural descriptors as influential variables; candidate transitions were interpreted as dataset-specific rather than universal thresholds or causal relationships. Three blind-well cases provided supplementary evidence of cross-well applicability, although larger independent multi-basin datasets are required to assess transferability.

Academic Editor: Tianshou Ma

Received: 8 June 2026

Revised: 21 July 2026

Accepted: 28 July 2026

Published: 31 July 2026

Keywords: deep shale gas; water saturation; physics-guided machine learning; ensemble learning; non-Archie conductivity; SHAP; Sichuan Basin

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

The deep Silurian shale gas reservoirs of the Sichuan Basin (burial depth > 3500 m) constitute a critical domain for reserve growth and production enhancement, yet their high water saturation (typically >40%) significantly constrains effective gas deliverability.

Accurate prediction of water saturation remains challenging because of the coexistence of low-resistivity effects and complex pore-scale water occurrence. Pyrite enrichment and organic-matter graphitization can markedly reduce formation resistivity, leading to systematic overestimation of water saturation by more than 20% using conventional Archie-based models [1]. At the pore scale, strong surface forces govern the distribution and thickness of water films in nanopores, while clay mineral transformations and surface hydration further complicate the distinction between bound and movable water [2]. In structurally deformed zones, fracture development further enhances water mobility, especially in high-curvature regions [3]. Recent studies of coal-bearing unconventional reservoirs have further demonstrated that fluid occurrence is closely coupled with surface wettability and mechanical response: Al_2O_3 nanofluids under different pH conditions and acidic SiO_2 nanofluids can alter solid–water contact angles, interfacial adsorption layers, mineral-surface reactions, and micromechanical properties [4,5]. At a larger scale, high-voltage electrical-pulse treatment has been shown to expand bottleneck pores, promote the transformation of closed pores into semi-open and open pores, generate connected fractures, and thereby enhance permeability and cross-scale gas transport [6]. Although coal and shale differ in mineral composition, organic-matter structure, and mechanical properties, these studies collectively indicate that wettability, pore-scale fluid occurrence, mechanical deformation, pore–fracture connectivity, permeability evolution, and gas migration are strongly coupled in unconventional reservoirs. These coupled effects render water saturation prediction in deep shale reservoirs a persistent challenge [7,8].

Traditional water saturation evaluation methods, centered on resistivity-based models, are fundamentally limited in such complex systems. The Archie equation assumes a homogeneous pore structure and neglects additional conductive pathways, restricting its applicability in clay-rich and organic-rich shales [9]. Extensions such as the Waxman–Smits model [10] and the dual-water model [11] partially account for clay conductivity and bound water effects, yet still rely on uncertain parameters and simplified assumptions. Even recent improvements, including modified Archie formulations [12] and J-function-based spatial modeling [13], fail to fully capture the coupled impacts of nanopore structure, fracture systems, and organic-matter evolution. Although advanced petrophysical techniques such as NMR provide direct insight into fluid occurrence and significantly improve saturation estimates [14–19], their high cost and limited field applicability restrict large-scale deployment. Consequently, no single method currently achieves both high accuracy and broad applicability in shale reservoirs [20,21].

Machine learning (ML) offers a promising alternative by leveraging multi-source data to model complex nonlinear relationships. Applications using well logs, core data, and seismic attributes have demonstrated substantial improvements in prediction accuracy over empirical models [22–24]. For example, ML-based approaches reduce prediction error without reliance on extensive core measurements [25] and perform well in low-resistivity, heterogeneous formations such as the Bakken [26] and tight sandstones [27]. However, purely data-driven models remain limited by their “black-box” nature, weak physical interpretability, and poor generalization in data-scarce environments. Hybrid approaches that integrate physical constraints with ML have therefore emerged, incorporating mechanisms such as DLVO-based water-film thickness or embedding governing equations via physics-informed neural networks [28,29]. Yet these approaches often remain “gray-box” models, with insufficient coupling between physical mechanisms and data-driven components. Despite this progress, four important gaps remain in current water saturation modeling. First, most machine-learning studies rely on raw well-log or empirically selected attributes, such that physical knowledge is invoked mainly for post hoc interpretation rather than being encoded in the model input. Second, existing hybrid approaches generally focus on a single mechanism, such as DLVO-controlled water-film thickness or governing-equation

regularization, and do not jointly address clay-related non-Archie conductivity and the occurrence state of pore water. Third, few models explicitly distinguish the controls on clay-bound water, organic-pore water, and fracture-connected movable water through computable descriptors coupling interfacial chemistry, pore connectivity, organic matter, and tectonic stress. Fourth, model performance is commonly assessed using random sample partitions within a single reservoir or region, leaving transfer across wells or blocks with different mineral compositions, salinities, and structural settings insufficiently tested. A unified framework is therefore required to embed multiple physical mechanisms before model learning, retain nonlinear predictive capability, and evaluate whether the learned relationships remain physically consistent and spatially transferable.

Accordingly, this study makes four main contributions. First, rather than treating water saturation as a direct function of resistivity or raw logging variables, we construct five groups of physics-derived descriptors representing water-film thickness, clay-related low-resistivity correction, pore connectivity, organic-pore contribution, and structural-stress effects. These descriptors provide a feature-level representation of the mechanisms controlling bound- and movable-water occurrence. Second, the physics-derived descriptors are integrated within a weighted RF–XGB–BNN ensemble that combines the robustness of random forests, the nonlinear learning capability of gradient boosting, and the flexible representation capability of the neural network component. The heterogeneous ensemble enables complementary prediction patterns from the three learners to be integrated without attributing the final performance to any single model architecture. Third, SHAP analysis is employed not merely to rank input variables, but to examine whether the effects and interactions learned by the model agree with expected geological and petrophysical mechanisms. Fourth, the framework is evaluated using multi-source data from several shale gas areas and independent blind wells to assess its spatial transferability. Together, these developments distinguish the proposed framework from conventional Archie-type models, purely data-driven machine learning, and single-mechanism physics–machine learning hybrids.

2. Materials and Methods

The overall framework of the proposed physics-guided machine-learning model is illustrated in Figure 1 and comprises two integrated components: a physics-based constraint module and a machine-learning prediction module. The physics module ingests multi-source data—including conventional well logs, core measurements, and seismic attributes—and transforms them into physically interpretable descriptors through submodules such as water-film thickness estimation, low-resistivity correction, and pore connectivity quantification. These features explicitly encode the underlying geological and petrophysical mechanisms. The machine-learning module then fuses the physics-derived descriptors with the original input features and performs feature engineering to construct an enriched feature space. The resulting feature set is subsequently fed into a weighted ensemble model that integrates random forest (RF), XGBoost (XGB), and a Bayesian neural network (BNN), as shown in Figure 1. This ensemble framework enables robust learning of complex nonlinear relationships, and its output is used as the final prediction of water saturation.

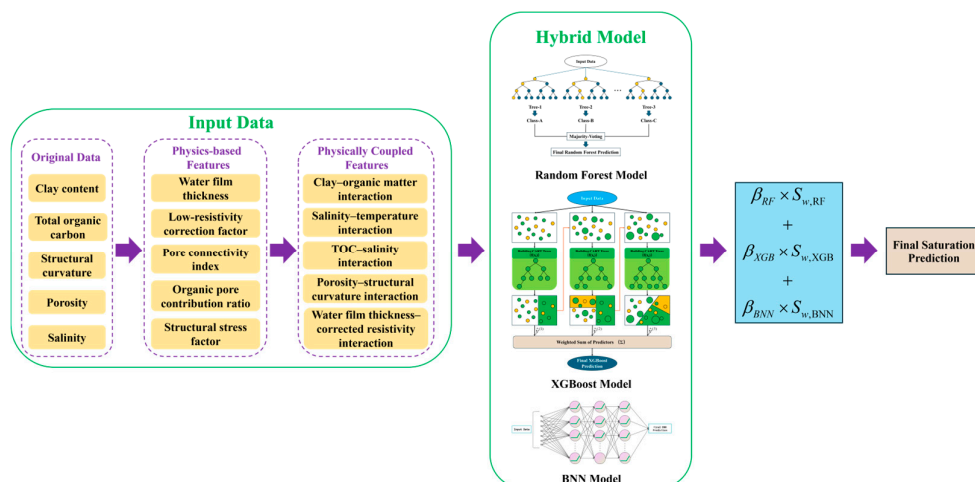


Figure 1. Architecture of the physics-guided machine-learning framework.

2.1. Physical Model

Building on the systematic analysis in Section 1 of the mechanisms controlling water saturation in deep shale gas reservoirs, this study extracts and constructs five core physics-based information modules from the key physical processes of interfacial chemistry, electrical response, pore structure, organic-matter evolution, and tectonic stress. These modules comprise a modified water-film thickness calculation model based on DLVO theory, a low-resistivity correction factor reflecting the conductive effect of organic-matter graphitization, a pore connectivity index characterizing the effective flow capacity of the pore network, an organic-pore contribution ratio quantifying the contribution of organic pores to reservoir storage space, and a tectonic-stress factor describing the intensity of structural deformation. This modular system explicitly represents multiple physical mechanisms in a computable form, enabling the embedding of physical constraints into the subsequent machine-learning framework and thereby improving the physical consistency and generalization capability of water saturation predictions.

It should be clarified that the five derived descriptors introduced below are semi-empirical formulations rather than strictly derived first-principles constitutive equations. Their functional forms were established according to the known qualitative relationships among interfacial chemistry, electrical conductivity, pore structure, organic-matter occurrence, structural deformation, and water saturation. The associated coefficients were empirically specified on the basis of domain knowledge, previously reported trends, and the parameter ranges represented by the present dataset. These coefficients were fixed before machine-learning training and were not optimized using the validation or test targets. Therefore, they should be interpreted as study-area-specific empirical parameters rather than universal physical constants. All input variables were entered using the units defined below, and the empirical coefficients implicitly contain the corresponding conversion dimensions required by each formulation. The applicability of these descriptors is limited to the geological conditions and parameter ranges covered by the studied samples; application to other reservoirs may require recalibration.

2.1.1. Water-Film Thickness

Clay content, formation-water salinity, temperature, and pressure jointly influence the occurrence of bound water on shale mineral surfaces. To represent their combined effects, a semi-empirical water-film descriptor was constructed using a DLVO-inspired functional form. This descriptor was developed for physics-guided feature engineering and was not

intended to provide a direct first-principles calculation or experimental measurement of absolute water-film thickness. It is expressed as follows:

$$\delta = \sqrt{\frac{A_{132}}{6\pi\kappa^{-1}\rho_w}} \quad (1)$$

$$A_{132} = 1.2 \times 10^{-20} + 2 \times 10^{-20} \phi_{\text{clay}} + 5 \times 10^{-21} \phi_{\text{clay}}^2 \quad (2)$$

$$\kappa^{-1} = \sqrt{\frac{\epsilon_r \epsilon_0 kT}{2 N_A e^2 I}} \approx \frac{0.304}{\sqrt{S}} \quad (3)$$

$$I = \frac{1}{2} \sum c_i z_i^2 \quad (4)$$

where δ denotes the dimensionless semi-empirical water-film development coefficient, which characterizes the relative tendency of water-film formation on clay mineral surfaces rather than the absolute physical film thickness. A_{132} is an effective empirical surface-interaction term for the clay–water–mineral three-phase system and characterizes the relative magnitude of interfacial van der Waals interactions (dimensionless). κ is the inverse Debye length (m^{-1}), whereas κ^{-1} represents the characteristic thickness of the diffuse electrical double layer (m). ρ_w is the water density (kg/m^3), and ϕ_{clay} is the dimensionless clay volume fraction expressed on a 0–1 scale. ϵ_r is the relative permittivity of water (dimensionless); ϵ_0 is the vacuum permittivity, $8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$; k_B is the Boltzmann constant, $1.38 \times 10^{-23} \text{ J/K}$; T is the absolute temperature (K); N_A is Avogadro's constant, $6.022 \times 10^{23} \text{ mol}^{-1}$; and e is the elementary charge, approximately $1.602 \times 10^{-19} \text{ C}$. I is the ionic strength in mol/m^3 ; c_i is the molar concentration of ionic species i in mol/m^3 ; and z_i is its valence (dimensionless). When the measured ionic concentrations are reported in mol/L , they are converted to mol/m^3 before calculating κ . S denotes formation-water salinity (g/L), reflecting the total concentration of dissolved salts in the pore water, and P denotes reservoir pressure (MPa). The quadratic clay-content term is introduced to represent the nonlinear influence of clay minerals on surface interactions and water-film development. The final water-film relationship is used as a unit-specific semi-empirical numerical-value equation, and its fitted coefficients incorporate the necessary unit normalization. Consequently, δ is reported as a dimensionless descriptor and is not interpreted as an absolute film thickness. The calculated descriptor was standardized together with the other predictors, and only its relative variation among samples was used in the machine-learning model. The adopted applicability range for this descriptor comprises a clay volume fraction of 0.15–0.45 (15–45%), formation-water salinity of 10–50 g/L , NaCl-equivalent ionic strength of approximately 0.17–0.86 mol/L (170–860 mol/m^3), and temperature of 353.15–425.15 K (80–152 °C).

2.1.2. Low-Resistivity Correction Factor

Clay minerals exhibit significant additional conductivity, mainly arising from ionic conduction pathways formed by exchangeable cations on their surfaces, which causes the formation resistivity (R_t) obtained from well logs to be systematically lower than the true resistivity—an effect that is particularly pronounced under conditions of high clay content or low salinity. To correct for this bias, this study constructs a semi-empirical correction factor inspired by the Waxman–Smits model to represent the influence of clay cation exchange capacity (CEC) on resistivity. The corrected resistivity can be expressed as:

$$R_t' = R_t \times (1.2 + 0.15 \text{CEC} - 0.003S + 0.002 \cdot \phi_{\text{clay}} \cdot \text{CEC}) \quad (5)$$

where R_t' denotes the corrected formation resistivity ($\Omega \cdot \text{m}$), and R_t is the original formation resistivity obtained from well logs ($\Omega \cdot \text{m}$). The term in parentheses represents the

dimensionless low-resistivity correction factor. Equation (5) is a unit-specific empirical numerical-value equation. In this equation, CEC , S , and ϕ_{clay} denote the numerical values of the clay cation exchange capacity expressed in meq/100 g, formation-water salinity expressed in g/L, and clay content expressed in %, respectively. For example, a clay content of 30% is entered as 30 rather than 0.30. Before calculation, all input variables were converted to these prescribed units. Consequently, the numerical constants 1.2, 0.15, 0.003, and 0.002 are treated as dimensionless empirical coefficients, and every term inside the parentheses is dimensionless. The corrected resistivity R_t' therefore retains the same unit as R_t , namely $\Omega\cdot\text{m}$. The empirical relationship is applicable only when the input variables are expressed using the prescribed units.

By introducing the clay cation exchange capacity, Equation (5) enhances the ability to characterize the additional conductivity associated with clay minerals. The interaction term between clay content and clay cation exchange capacity further describes the conductive contribution arising from the combined effects of clay type and content. The salinity term reflects the controlling effect of high-salinity fluids on the increase in electrical conductivity; that is, the higher the salinity, the stronger the solution conductivity, the lower the measured R_t , and the stronger the correction required. The adopted applicability range for this correction factor is an original formation resistivity of 1–30 $\Omega\cdot\text{m}$, clay cation exchange capacity of 5–25 meq/100 g, formation-water salinity of 10–50 g/L, and clay content of 15–45%.

2.1.3. Pore Connectivity Index

To represent the relative tendency of pore–fracture connectivity of the shale pore–fracture system, this study constructs a pore connectivity index that jointly accounts for the promoting effect of structural curvature on microfracture development and the suppressing effect of clay minerals on pore-throat blockage. The pore connectivity index can be expressed as:

$$CI = [0.5 \times \text{sigmoid}(K_c) + 0.3] \times (1 - \phi_{\text{clay}}) \quad (6)$$

where CI denotes the dimensionless pore connectivity index. Equation (6) is a unit-specific semi-empirical numerical-value equation. In this equation, K_c denotes the numerical value of structural curvature expressed in km^{-1} . For example, a structural curvature of 0.08 km^{-1} is entered as 0.8. Therefore, K_c is treated as a dimensionless numerical input within the sigmoid function. ϕ_{clay} denotes the dimensionless clay volume fraction expressed on a 0–1 scale; for example, a clay content of 30% is entered as 0.30. Accordingly, the clay fraction used in Equation (6) can be obtained from the percentage clay content as $\phi_{\text{clay}} = C_{\text{clay},\%}/100$. Before calculation, structural curvature and clay content were converted to these prescribed forms.

The sigmoid function is defined as:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (7)$$

where x is a dimensionless numerical input. In Equation (6), x corresponds to the numerical value of the structural curvature expressed in km^{-1} . Consequently, the exponent $-x$ is dimensionless. The constants 0.5 and 0.3 are also dimensionless empirical coefficients, and both $\text{sigmoid}(K_c)$ and $1 - \phi_{\text{clay}}$ are dimensionless. Therefore, the resulting CI is dimensionless.

The sigmoid function maps the prescribed curvature value to the interval between 0 and 1 and represents a smooth nonlinear increase in the relative contribution of curvature-induced fractures. The term $1 - \phi_{\text{clay}}$ describes the suppressive influence of clay enrichment on pore connectivity. Higher clay content generally increases pore-throat filling and blockage, thereby restricting the effective pathways available for movable fluids. In contrast, increasing structural curvature is associated with a greater tendency for microfracture development and enhanced pore–fracture connectivity. Because the coefficients were

empirically specified, CI represents a relative connectivity descriptor rather than a direct measurement of absolute pore-network connectivity. Its interpretation and application are limited to the structural-curvature and clay-content ranges represented by the calibration dataset. The adopted applicability range for this index is a structural curvature of 0.01–0.10 km^{-1} and a clay content of 15–45%, with the clay content entered into Equation (6) as a dimensionless volume fraction of 0.15–0.45.

2.1.4. Formulation of the Organic-Pore Contribution Ratio

To represent the relative contribution tendency of organic matter to the effective storage space under different mineralogical backgrounds, this study develops an organic-pore contribution ratio model. This model comprehensively accounts for the coupled relationships among total organic carbon content, total porosity, and clay mineral content, and is expressed as:

$$\phi_{org} = \frac{0.25 \times TOC \times \phi}{1 + 0.1 \times \phi_{clay}} \quad (8)$$

where ϕ_{org} denotes the dimensionless semi-empirical organic-pore contribution index rather than a directly measured organic-pore volume fraction. TOC , ϕ , and ϕ_{clay} denote the dimensionless total organic carbon fraction, total porosity, and clay fraction, respectively, and are entered on a 0–1 scale. For example, a TOC content of 4%, a total porosity of 5%, and a clay content of 30% are entered as 0.04, 0.05, and 0.30, respectively. Before calculation, all percentage measurements are divided by 100 and converted to dimensionless fractions.

The coefficients 0.25 and 0.1 are dimensionless empirical coefficients. Therefore, the numerator $0.25 \times TOC \times \phi$ is dimensionless, and the denominator $1 + 0.1 \times \phi_{clay}$ is also dimensionless. Consequently, the resulting ϕ_{org} is dimensionless, and Equation (8) is dimensionally consistent.

The numerator represents the combined positive influence of organic-matter abundance and total pore volume on the relative development of organic pores. The denominator introduces an empirical clay-adjustment term to account for the suppressive influence of clay enrichment on the effective contribution of organic pores. Clay minerals may occupy pore space, increase surface adsorption and capillary confinement, and reduce the effective connectivity of organic pores. Accordingly, increasing TOC and porosity increases the index, whereas increasing clay content decreases it.

Because the coefficient values were empirically specified, ϕ_{org} should be interpreted as a relative organic-pore contribution descriptor rather than an absolute organic-pore volume fraction. The calculated descriptor was standardized together with the other predictors before model training, and only its relative variation among samples was used by the machine-learning model. Its application should therefore be restricted to the variable ranges represented by the calibration dataset. The adopted applicability range for this index is a TOC content of 1.5–6.5%, total porosity of 3–12%, and clay content of 15–45%; all percentage variables are divided by 100 and entered into Equation (8) as dimensionless fractions.

2.1.5. Formulation of the Structural-Stress Factor

Structural deformation can affect the preservation, compaction, and connectivity of shale pore systems, particularly in deep reservoirs where pore structures are sensitive to regional bending and deformation. To describe the relative structural condition of the reservoir, this study constructs a semi-empirical structural-stress proxy based on structural curvature:

$$S_f = \exp \left[-0.5 \times \left(\frac{K_c - 0.06}{0.02} \right)^2 \right] \quad (9)$$

where S_f denotes the dimensionless structural-stress factor. Equation (9) is a unit-specific semi-empirical numerical-value equation. In this equation, K_c represents the numerical value of structural curvature expressed in km^{-1} . For example, a structural curvature of 0.06 km^{-1} is entered as 0.06. Before calculation, all curvature measurements were converted to this prescribed unit. The values 0.06 and 0.02 represent, respectively, the empirical reference curvature and the characteristic response width in the prescribed curvature unit, while 0.5 is a dimensionless empirical coefficient. Therefore, $(K_c - 0.06)/0.02$ is dimensionless, the exponent is dimensionless, and the resulting S_f is also dimensionless.

The Gaussian-type response reaches its maximum value of 1 when $K_c = 0.06 \text{ km}^{-1}$ and decreases as the curvature deviates from this reference value. The reference value of 0.06 km^{-1} represents an empirically selected central structural condition within the calibration dataset, whereas 0.02 km^{-1} controls the sensitivity of S_f to curvature deviation. These values are empirical calibration parameters rather than universal mechanical thresholds.

This formulation represents the assumption that moderate structural deformation may favor the preservation and effective connectivity of the pore–fracture system, whereas excessively weak or strong deformation may be associated with insufficient fracture development, pore compression, structural damage, or reduced connectivity. Because S_f is derived solely from structural curvature, it should be interpreted as a relative structural-stress proxy rather than a direct measurement of in situ stress magnitude. The calculated factor was standardized together with the other predictors before model training, and only its relative variation among samples was used by the machine-learning model. Its application should therefore be restricted to the curvature range represented by the calibration dataset. The adopted applicability range for this factor is a structural curvature of $0.01\text{--}0.10 \text{ km}^{-1}$, with the strongest response concentrated near $0.04\text{--}0.08 \text{ km}^{-1}$.

2.2. Machine-Learning Model

2.2.1. Feature-Engineering Module

To enhance the model's ability to characterize the complex reservoir mechanisms of deep shale gas, this study constructs interaction features with clear physical meaning on the basis of the five categories of physical factors, so as to capture the coupled relationships among key geological parameters. For example, the synergistic effect between clay and organic matter is represented by $\phi_{\text{clay}} \times \text{TOC}/100$, which reflects the competing and promoting effects between clay adsorption properties and organic-pore development; the coupling between salinity and temperature is described by $S \times T/1000$, capturing the role of elevated temperature in enhancing ionic activity and thus modifying the aqueous geochemical environment; the interaction term $\text{TOC} \times S/100$ between organic matter and salinity is used to characterize the joint control of organic-pore adsorption capacity and formation-water salinity; the response of porosity to structural curvature is expressed quantitatively as $\phi \times K_c \times 100$, thereby making explicit the impact of structural-stress perturbations on pore connectivity; and the coupled feature $\delta \times R'_f/100$ between water-film thickness and corrected resistivity reveals the linkages among water-film control and bound-water proportion, conductive pathways, and resistivity response. In addition, to strengthen the model's capability to identify higher-order nonlinear coupling relationships, the feature space is further expanded using a second-order polynomial scheme, enabling a more comprehensive representation of the complex interactions among structure, mineralogy, aqueous geochemistry, and pore structure.

2.2.2. Model Architecture and Training

The final prediction was generated using a weighted ensemble of random forest (RF), XGBoost (XGB), and a Bayesian neural network (BNN). All continuous predictors were standardized using the means and standard deviations calculated exclusively from the training set,

and the same transformation parameters were subsequently applied to the validation and test sets. Model hyperparameters were determined by five-fold cross-validation within the training set, using the mean absolute error (MAE) as the optimization criterion. The independent validation set was used to determine the ensemble weights and perform early stopping, whereas the test and blind-well datasets were excluded from model selection.

The RF model consisted of 300 regression trees, with a maximum tree depth of 8, a minimum of 5 samples required for internal-node splitting, and a minimum leaf size of 1. Bootstrap sampling and the squared-error criterion were used. The XGB model consisted of 300 boosting trees, with a maximum depth of 6, a learning rate of 0.01, a subsampling ratio of 0.8, and a column-sampling ratio of 1.0. The regression squared-error objective was adopted. The random seed was fixed at 42 for all models.

The BNN contained three fully connected hidden layers with 128, 64, and 32 neurons, respectively, followed by a single linear regression-output node. Rectified linear unit functions were used in the hidden layers. Independent standard Gaussian priors,

$$p(\theta) = \mathcal{N}(0, 1), \quad (10)$$

were assigned to the network weights and biases. A mean-field Gaussian distribution was adopted as the variational posterior:

$$q(\theta) = \mathcal{N}(\mu, \sigma^2), \quad (11)$$

where μ and σ were optimized using the reparameterization method. Variational inference was performed by minimizing the negative evidence lower bound, which combines the expected negative log-likelihood and the Kullback–Leibler divergence between the variational posterior and the prior. The BNN was trained using the Adam optimizer, with an initial learning rate of 1×10^{-3} , a batch size of 32, and a maximum of 500 epochs. Training was terminated early when the validation MAE did not improve for 25 consecutive epochs.

The validation MAEs of RF, XGB, and BNN were 4.9%, 4.6%, and 4.8%, respectively. These validation errors were first converted into normalized performance scores using the reciprocal–MAE relationship:

$$\tilde{w}_m = \frac{\text{MAE}_m^{-1}}{\sum_{j=1}^3 \text{MAE}_j^{-1}} \quad (12)$$

where m represents RF, XGB, or BNN. The resulting initial weights were 0.324, 0.345, and 0.331, respectively. Because the optimal ensemble weights also depended on the complementarity and correlation among the residuals of the three models, these initial weights were further refined on the validation set. Non-negative weight combinations centered on the initial values were evaluated at an interval of 0.05 under the constraint

$$w_{\text{RF}} + w_{\text{XGB}} + w_{\text{BNN}} = 1 \quad (13)$$

The optimal combination was determined by minimizing the validation-set ensemble MAE:

$$(w_{\text{RF}}^*, w_{\text{XGB}}^*, w_{\text{BNN}}^*) = \arg \min_w \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \left| y_i - \sum_m w_m \hat{y}_{i,m} \right| \quad (14)$$

The weight combination of 0.35:0.35:0.30 produced the lowest ensemble MAE on the validation set and was therefore fixed before evaluation on the test and blind-well datasets.

During inference, 100 Monte Carlo samples were drawn from the learned variational posterior for each input, and the mean of the sampled predictions was used as the BNN point prediction. The posterior dispersion was not used to construct calibrated sample-level prediction intervals in the present study. Accordingly, the BNN component was employed primarily as a probabilistically regularized nonlinear learner within the heterogeneous

ensemble rather than as a standalone uncertainty-quantification model. The final ensemble prediction was calculated as

$$\hat{S}_w = 0.35\hat{S}_{w,RF} + 0.35\hat{S}_{w,XGB} + 0.30\hat{S}_{w,BNN} \quad (15)$$

The main hyperparameters and training settings of the RF, XGB, and BNN models are summarized in Table 1. These parameters were determined through five-fold cross-validation within the training set using MAE as the optimization criterion and were subsequently fixed during validation, test-set evaluation, and blind-well prediction.

Table 1. Main hyperparameters and training settings of the RF, XGB, and BNN models.

Model	Architecture/Size	Optimization and Training	Regularization and Inference
RF	300 trees; maximum depth = 8	Minimum split size = 5; minimum leaf size = 1	Squared-error criterion; bootstrap sampling
XGB	300 trees; maximum depth = 6	Learning rate = 0.01	Squared-error objective; subsample = 0.8; column sample = 1.0
BNN	Fully connected 128–64–32–1 network; ReLU activation	Adam optimizer; learning rate = 1×10^{-3} ; batch size = 32; maximum epochs = 500; early-stopping patience = 25	$\mathcal{N}(0,1)$ parameter prior; mean-field Gaussian posterior; Monte Carlo mean used for point prediction

Note: The random seed was fixed at 42 for all models. The reported hyperparameters were determined using five-fold cross-validation within the training set and were fixed before validation, test-set evaluation, and blind-well prediction.

2.2.3. Final Water Saturation Prediction Equation

To achieve high-accuracy prediction of water saturation, this study deeply integrates the physics-based mechanism modules with the data-driven model and constructs a hybrid prediction framework with explicit physical meaning. The water saturation S_w can be written as a composite function of the physical feature mapping and the ensemble learning model:

$$S_w = f_{ensemble} \{ \delta(R_t, S, Clay), R'_t(R_t, CEC, S), CI(Clay, K_c), \phi_{org}(TOC, \phi, Clay), S_f(K_c), TOC, \phi, K_c, \phi_{clay} \} \quad (16)$$

where the physical features correspond, respectively, to the clay-surface water-film thickness δ , the resistivity R'_t corrected for clay-related conductivity, the pore connectivity index CI , the organic-pore contribution ratio ϕ_{org} , and the structural-stress factor S_f , among other key physical mechanisms. Through this explicit construction, the model can fully preserve the controlling effects of physical processes on water saturation, while providing the machine-learning component with structured and interpretable input features.

On the data-driven side, the final prediction is generated by a weighted ensemble model composed of random forest, XGBoost, and the Bayesian neural network. The ensemble output can be expressed as:

$$S_w = 0.35 \times S_{w,RF} + 0.35 \times S_{w,XGB} + 0.30 \times S_{w,BNN} \quad (17)$$

The individual weights were determined according to the relative MAE performance on the validation set to achieve a balance between predictive accuracy and robustness. The weighted ensemble combined the complementary prediction patterns of RF, XGB, and BNN, while no quantitative claim regarding calibrated predictive uncertainty was made in the present study. By explicitly embedding physical mechanisms into the input space and employing ensemble learning to enhance algorithmic performance, this framework effectively overcomes the limitations of traditional physics-based models, as well as the strong data

dependence and insufficient physical interpretability inherent in purely data-driven machine-learning approaches.

2.2.4. Model Evaluation and Robustness Analysis

Model performance was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). MAE and RMSE were expressed in percentage points of water saturation. Before model development, the 153 samples were divided into a training set of 107 samples, a validation set of 16 samples, and an independent test set of 30 samples using target-stratified random sampling.

Five-fold cross-validation was conducted exclusively within the 107-sample training set to evaluate model stability. The independent validation set was used for hyperparameter selection and ensemble-weight determination, whereas the independent test set remained completely excluded from model development and was evaluated only once after the model configuration had been finalized. Within each cross-validation fold, missing-value imputation, outlier treatment, normalization, feature construction, feature selection, and model fitting were performed using only the corresponding fold-training subset.

2.3. Dataset Description and Preprocessing

2.3.1. Dataset Composition and Multi-Source Data Integration

A total of 153 depth-matched samples from five wells in the Dingshan area were used for model development and evaluation. The dataset contained 27, 5, 47, 35, and 39 samples from wells DY1, DY2, DY3, DY4, and DY5, respectively, covering a depth range of 1928.01–4366.54 m. Each record was assigned a unique identifier consisting of the well name, sample identifier, and sampling depth to ensure traceability throughout data preprocessing and model evaluation.

The dataset integrated laboratory measurements, core-derived petrophysical properties, conventional well-log responses, and reservoir-condition parameters defined at the sample, well, or study-area scale. Measured water saturation was used as the prediction target. TOC, clay content, porosity, and permeability were obtained from the corresponding sample or the nearest available core-analysis record. A maximum depth difference of 1.0 m was allowed when matching laboratory and core-derived variables. Acoustic transit time (AC), density (DEN), gamma ray (GR), and resistivity (R_t) were matched to the sample depths using a maximum depth difference of 0.125 m. Temperature, pressure, pressure coefficient, and other reservoir-condition parameters were assigned according to their applicable spatial resolution. Before integration, all variables were converted to consistent units and referenced to a unified depth system. The ranges, descriptive statistics, and data completeness of the original variables are summarized in Table 2.

Table 2. Descriptive statistics of the input variables and measured water saturation for the 153 samples.

Variable	Range	Mean $\pm$ SD/Reference Value	Median
Depth (m)	1928.01–4366.54	2981.91 $\pm$ 822.97	3587.75
TOC (%)	0.20–6.67	1.51 $\pm$ 1.28	0.97
Clay content (%)	14.90–62.10	43.15 $\pm$ 10.99	45.65
CEC (meq/100 g)	12.3–18.9	15.60 $\pm$ 1.65	15.6
Salinity (g/L)	24.2–36.7	30.45 $\pm$ 3.13	30.45
Porosity (%)	1.30–7.32	3.11 $\pm$ 1.20	2.86
Permeability (mD)	0.00012–144.97590	9.653 $\pm$ 30.205	0.0651
R_t ($\Omega \cdot m$)	8.93–395.66	71.44 $\pm$ 63.53	49.36
K_c (km^{-1})	0.041–0.068	0.0545 $\pm$ 0.0068	0.0545

Temperature (°C)	53.00–66.00	59.48 ± 5.16	61.00
Pressure (MPa)	20.13–65.08	39.70 ± 16.46	54.10
Pressure coefficient (–)	0.98–1.49	1.25 ± 0.21	1.45
AC (µs/ft)	66.22–86.49	75.70 ± 3.84	75.69
DEN (g/cm ³)	2.42–2.74	2.61 ± 0.07	2.63
GR (API)	94.82–492.36	164.36 ± 54.20	150.69
Measured S_w (%)	12.46–92.94	41.44 ± 12.14	40.60

2.3.2. Data Quality Control, Missing-Value Treatment, and Normalization

Data quality control was performed before model development. Samples without measured water saturation were excluded because water saturation was used as the prediction target. Records with duplicated well and depth identifiers were checked, and the consistency of variable names, measurement units, depth references, and physically plausible ranges was verified against the original data files. Values associated with unit inconsistencies, data-entry errors, or unsuccessful depth matching were removed, whereas geologically meaningful extreme values reflecting reservoir heterogeneity were retained.

For predictor variables containing missing values, missing entries were replaced using the median calculated exclusively from the corresponding training subset. The same training-derived medians were subsequently applied to the validation and test subsets. Binary missingness indicators were generated as candidate features, allowing the model to distinguish between originally observed and imputed values. Missing-value treatment was performed independently within each cross-validation fold and did not introduce additional synthetic samples.

After construction of the physics-derived and interaction descriptors, continuous candidate features were clipped to the 1st and 99th percentiles estimated from the corresponding training subset to reduce the influence of extreme numerical values. Features with a training-set variance below 10^{-12} were removed. The remaining continuous variables were standardized using

$$x_{\text{std}} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}, \quad (18)$$

where μ_{train} and σ_{train} represent the mean and standard deviation calculated from the training subset. The clipping thresholds and normalization parameters obtained from the training subset were applied unchanged to the corresponding validation and test subsets.

Feature selection was subsequently performed independently within each training subset, and the 20 highest-ranked candidate features were retained for model development. Imputation, outlier treatment, normalization, variance filtering, feature selection, hyperparameter optimization, and ensemble-weight determination were therefore completed without access to the corresponding test subset. This fold-specific preprocessing workflow reduced the risk of information leakage and ensured that the reported test performance reflected predictions for samples not used during model development.

2.3.3. Dataset Partitioning and Leakage Control

The complete dataset was randomly divided into a training set ($n = 107$), a validation set ($n = 16$), and an independent test set ($n = 30$). Five-fold cross-validation was conducted exclusively within the training set. The five held-out folds contained 22, 22, 21, 21, and 21 samples, respectively. The validation set was used for model selection and ensemble-weight determination, whereas the independent test set remained completely isolated until the final performance evaluation.

During each cross-validation fold, missing-value imputation, outlier treatment, normalization, variance filtering, feature construction, and feature selection were performed

using only the fold-specific training portion. The resulting preprocessing parameters were then applied unchanged to the corresponding held-out fold. This procedure prevented information from the held-out samples from entering model fitting or preprocessing.

After model selection, the preprocessing pipeline and candidate models were refitted using the complete training set of 107 samples. The fixed validation set was used to determine the final model configuration and ensemble weights, whereas the independent test set was accessed only once to obtain the final performance metrics. Consequently, the five-fold results represent internal cross-validation within the training set, while the 30-sample test results represent the independent sample-wise evaluation.

3. Results and Discussion

3.1. Model Performance and Generalization Assessment

3.1.1. Five-Fold Cross-Validation and Generalization Performance

The complete dataset was first divided into a fixed training set of 107 samples, a validation set of 16 samples, and an independent test set of 30 samples. Five-fold cross-validation was subsequently conducted exclusively within the training set to evaluate model stability and support model development. The five held-out cross-validation folds contained 22, 22, 21, 21, and 21 samples, respectively. The validation set was used for final model selection and ensemble-weight determination, whereas the independent test set remained completely isolated from preprocessing, feature selection, hyperparameter optimization, and ensemble-weight determination.

The five-fold cross-validation results showed that the R^2 values ranged from 0.894 to 0.919, while MAE and RMSE ranged from 1.55% to 1.83% and from 2.08% to 2.44%, respectively. The mean cross-validation R^2 , MAE, and RMSE were 0.907 ± 0.009 , $1.69\% \pm 0.10\%$, and $2.25\% \pm 0.14\%$, respectively. The relatively small fold-to-fold variations indicated that model performance was not highly sensitive to a particular partition of the training samples.

After model selection, the final model was fitted using the complete training set and evaluated on the fixed validation and independent test sets. The model achieved R^2 values of 0.955, 0.904, and 0.902 on the training, validation, and test sets, respectively. The corresponding MAEs were 0.98%, 1.76%, and 1.77%, while the RMSE values were 1.31%, 2.31%, and 2.34%, respectively. The training-to-validation differences were 0.051 for R^2 , 0.78 percentage points for MAE, and 1.00 percentage point for RMSE, indicating a moderate fitting advantage on the training data. However, the validation and independent test results were very similar, with differences of only 0.002 in R^2 , 0.01 percentage points in MAE, and 0.03 percentage points in RMSE. The independent test results were also close to the mean cross-validation performance, differing by 0.005 in R^2 , 0.08 percentage points in MAE, and 0.09 percentage points in RMSE. This agreement, together with the limited variation among the cross-validation folds, suggests that the model did not exhibit pronounced overfitting. Nevertheless, because the independent test set contained only 30 samples, these generalization results should be interpreted together with the cross-validation and blind-well evaluations. Figure 2 presents the prediction results for the 30-sample independent test set. As shown in Figure 2a, the predicted water saturation values generally follow the 1:1 reference line across the observed saturation range. Figure 2b shows that the prediction errors are distributed around zero without a pronounced directional bias. Because the test samples were isolated before preprocessing, feature selection, hyperparameter optimization, and ensemble-weight determination, these results provide an independent assessment of the model performance under the adopted sample-wise partition.

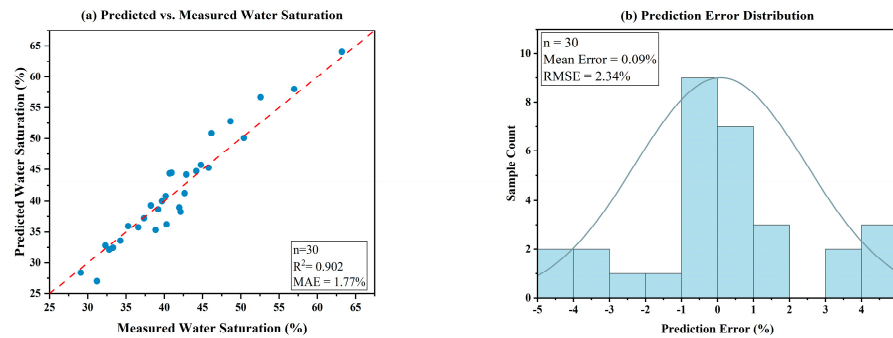


Figure 2. Generalization performance of the model on the independent test set: (a) predicted versus measured water saturation and (b) distribution of prediction errors.

3.1.2. Comparative Performance and Contribution of Physics-Derived Descriptors and Ensemble Strategy

To quantify the respective contributions of the physics-derived descriptors and the ensemble strategy, the proposed PhysML-Hybrid model was compared with the conventional Archie model, a purely data-driven ensemble (ML-only), a physics-guided XGBoost model (Phy-XGB), and a support vector machine (SVM) using identical data partitions and preprocessing procedures. The ML-only model retained the ensemble architecture but excluded the physics-derived descriptors, whereas Phy-XGB used the same physics-derived feature set as PhysML-Hybrid but employed XGBoost as the sole predictor. The SVM employed a radial basis function kernel and was included as a representative nonlinear machine-learning benchmark. These configurations enabled the effects of the physics-derived descriptors, ensemble strategy, and nonlinear prediction capability to be evaluated separately.

As illustrated in Figure 3, PhysML-Hybrid achieved the highest coefficient of determination and the lowest mean and maximum absolute errors among all evaluated models. Phy-XGB produced the second-best overall performance, followed by SVM and ML-only, whereas the conventional Archie model exhibited the largest prediction errors. The improvement over SVM indicates that the performance gain cannot be attributed solely to the use of a nonlinear predictor. Meanwhile, the improvements over ML-only and Phy-XGB demonstrate the respective contributions of the physics-derived descriptors and the weighted integration of RF, XGB, and BNN predictions. Collectively, these comparisons support the effectiveness of combining physical knowledge with a heterogeneous ensemble learning strategy.

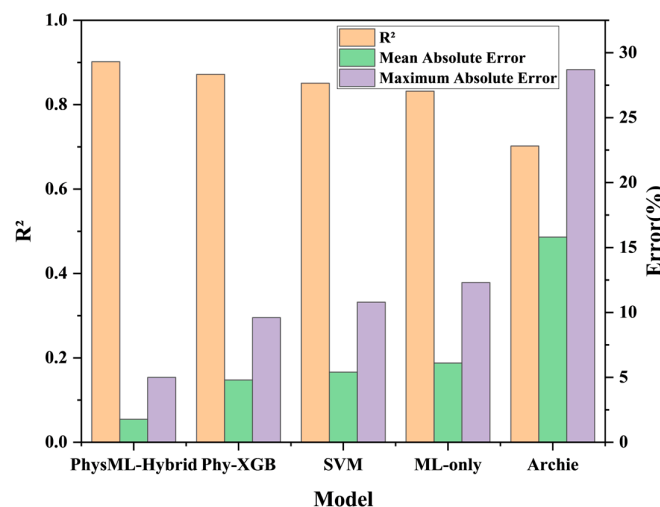


Figure 3. Prediction performance of different water saturation models.

3.1.3. Depth-Stratified Performance and Robustness Analysis

To further examine the depth-dependent robustness of PhysML-Hybrid, the pooled out-of-fold predictions were stratified into five depth intervals, as summarized in Table 3. The model exhibited its best performance within the 1900–2100 m interval, achieving an R^2 of 0.919, an MAE of 1.58%, and an RMSE of 2.08%. Comparable performance was obtained within the 2100–2300 m and 3500–3700 m intervals, where the R^2 values remained above 0.90. A moderate reduction in predictive accuracy was observed within the 3700–3900 m interval, with the R^2 decreasing to 0.888 and the MAE and RMSE increasing to 1.95% and 2.55%, respectively. The 4300–4400 m interval exhibited the largest errors, with an R^2 of 0.840, an MAE of 2.35%, and an RMSE of 2.95%. However, this deepest interval contained only five samples; therefore, its performance estimates should be interpreted cautiously. Overall, the depth-stratified results indicate that PhysML-Hybrid maintained relatively stable predictive performance across the principal sampled depth intervals, although accuracy decreased moderately in the deepest and most sparsely sampled interval.

To complement the interval-based evaluation in Table 3, Figure 4 presents a depth-aligned prediction profile for eight test samples from well DY4 over the interval of 3582.355–3712.325 m. The predicted water saturation profile generally follows the measured vertical variations, including the increase near 3652–3654 m and the subsequent decrease toward the lower part of the analyzed interval. Relatively larger local deviations occur around 3637–3652 m, where the clay content, porosity, and other petrophysical properties also vary markedly, suggesting that abrupt changes in reservoir properties remain more difficult to represent accurately. Nevertheless, no persistent vertical offset is observed between the measured and predicted profiles, indicating that the model can reproduce the principal depth-dependent variation in water saturation. Because permeability measurements were unavailable at these depths, the permeability track was obtained by KNN imputation and is included only to provide supplementary petrophysical context. Given the limited number of samples, this single-well profile should be regarded as supporting evidence for the depth-stratified results rather than as an independent validation dataset.

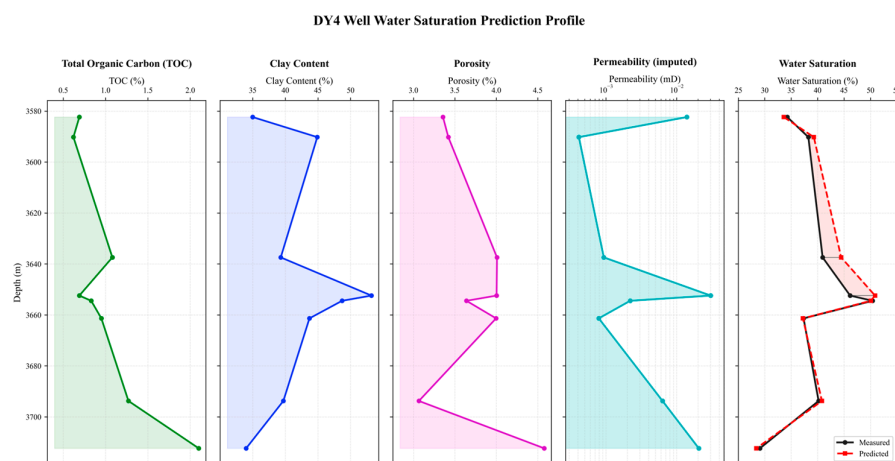


Figure 4. Depth-aligned petrophysical properties and measured and predicted water saturation profiles for eight test samples from well DY4.

Table 3. Depth-stratified performance of the PhysML-Hybrid model.

Depth Interval (m)	Sample Size	R^2	MAE (%)	RMSE (%)
1900–2100	27	0.919	1.58	2.08
2100–2300	47	0.913	1.66	2.20

3500–3700	26	0.904	1.74	2.33
3700–3900	48	0.888	1.95	2.55
4300–4400	5	0.840	2.35	2.95

3.1.4. Blind-Well Validation

To provide a preliminary assessment of the cross-well applicability of the model, three representative deep shale gas wells in southeastern Sichuan that were excluded from model development were selected for blind testing. Key inputs, including geological characteristics, petrophysical properties, and reservoir-engineering parameters, were compiled to perform independent water saturation prediction. The input parameters and water saturation prediction results for the three blind-well cases are summarized in Table 4.

The three blind-well cases yielded relative errors ranging from 3.09% to 9.97%, providing preliminary supplementary evidence of cross-well applicability. Given that only one representative sample was available for each well, these results should not be interpreted as validation of continuous vertical trends or definitive cross-block generalization.

Table 4. Input parameters and water saturation prediction results for the three blind-well cases.

Well Name	Structural Curvature	TOC (%)	Clay Minerals (%)	Salinity (g/L)	Resistivity ($\Omega\cdot\text{m}$)	Porosity (%)	Permeability (md)	Formation Temperature ($^{\circ}\text{C}$)	Formation Pressure (MPa)	Pressure Coefficient	Burial Depth (m)	Water Saturation (%)	Predicted Water Saturation (%)	Relative Error (%)
DY18	0.07	3.72	40.10	42.50	1.8	5.66	/	152	89	1.85	4794	56.00	60.31	7.70
DYS5	0.06	3.61	37.46	20.71	6	5.4	0.014	152	89	1.85	4844	46.89	48.34	3.09
QYS1	0.05	3.67	35.19	42.94	26.94	7.08	0.414	152	89	1.85	4812	37.73	41.49	9.97

3.2. Parameter Sensitivity Analysis

To elucidate the decision logic of the proposed model, SHapley Additive exPlanations (SHAPs) were used to quantify the relative contributions of the geological, petrophysical, and physics-derived input variables. As shown in Figure 5, clay content exhibited the largest mean absolute SHAP value, followed by the curvature–clay index, total organic carbon (TOC), pore connectivity index, and structural curvature. The dominant contribution of clay content indicates that the model predictions were particularly sensitive to clay-related water retention and electrical-conductivity effects. The high importance of the curvature–clay index further suggests that the model captured the combined influence of structural deformation and clay-rich lithology rather than treating these factors as entirely independent controls. The contributions of TOC, pore connectivity, and structural curvature indicate that organic-matter development, pore-network continuity, and structural deformation also provided important information for water saturation prediction. Water-film thickness, organic-pore contribution, and the remaining interaction and higher-order terms made smaller but nonzero contributions to the model output.

Because mean absolute SHAP values quantify the average magnitude of a variable's contribution rather than the direction of its effect, the ranking in Figure 5 alone cannot determine whether an increase in a given variable raises or lowers water saturation. To move beyond this global importance ranking and assess the nonlinear and threshold-like relationships suggested by the model, the five highest-ranked variables were further examined using all 153 processed samples (Figures 6–10). Spearman correlation analysis, LOESS regression with bootstrap 95% confidence intervals, sample-distribution analysis, and continuous piecewise linear regression were jointly employed to evaluate the direction, stability, and potential transition points of these relationships.

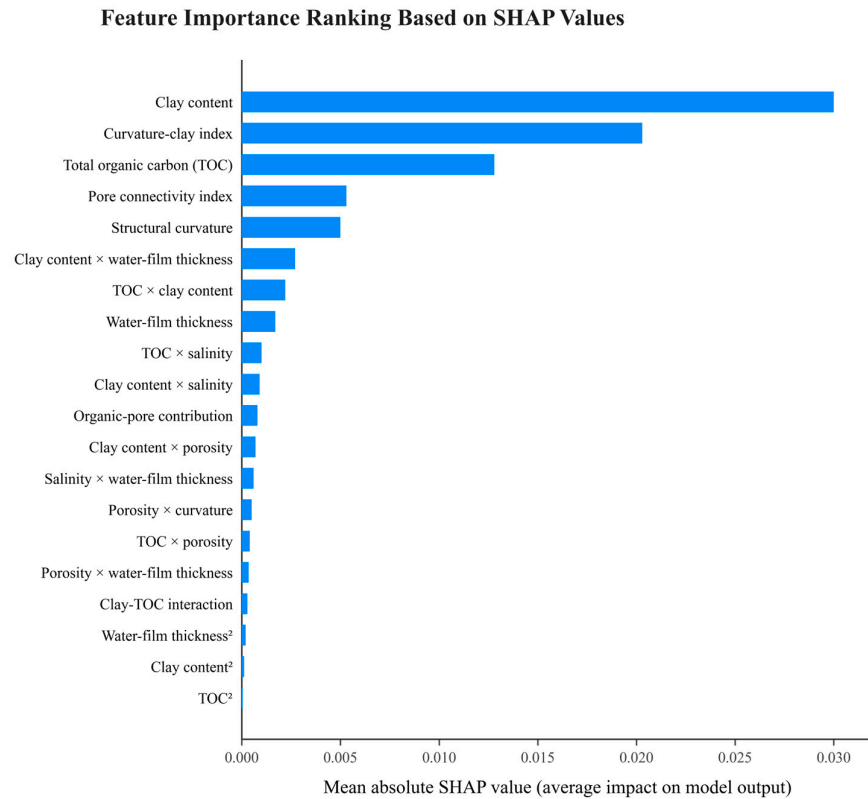


Figure 5. Feature-importance ranking of key sensitive parameters based on SHAP values.

The estimated breakpoints were evaluated in conjunction with their confidence intervals, local sample densities, and improvements in piecewise-regression fit. Breakpoints associated with wide confidence intervals, sparse observations, or only limited improvements in model fit were interpreted as dataset-specific candidate transitions rather than universal physical thresholds. Moreover, SHAP values characterize how the trained model uses individual variables and should not be regarded as direct evidence of geological causality. The geological interpretations presented below therefore combine the model-derived attribution patterns with sample-level statistical evidence and established petrophysical mechanisms.

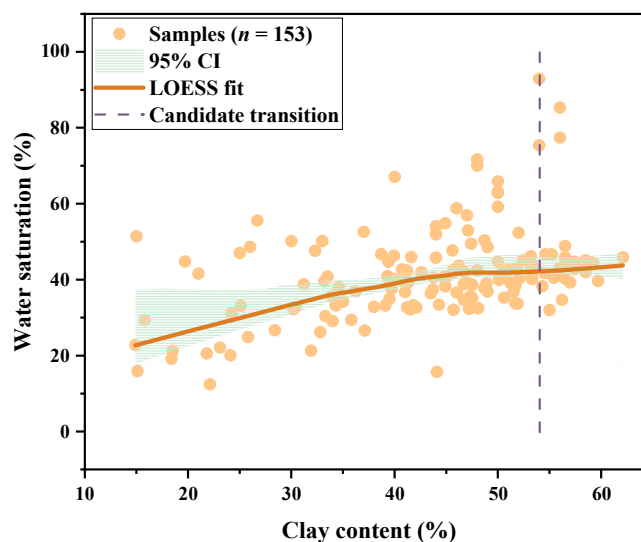


Figure 6. Relationship between clay content and water saturation.

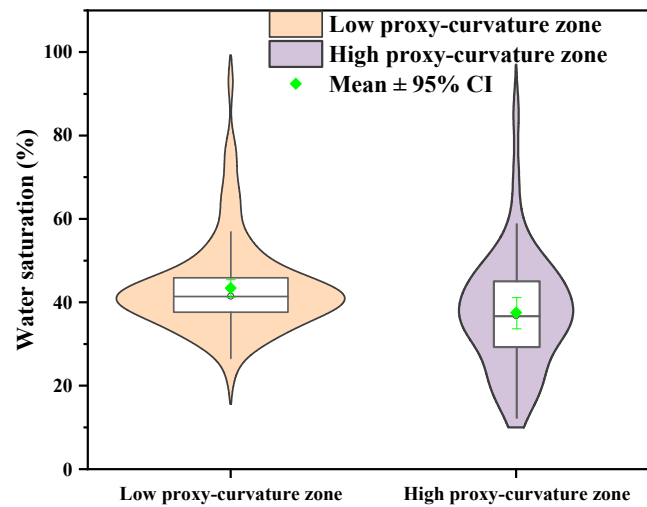


Figure 7. Relationship between structural curvature and water saturation.

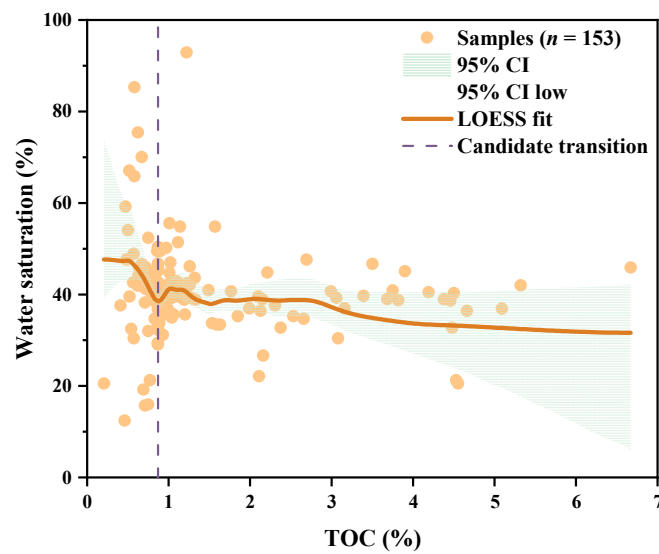


Figure 8. Relationship between TOC content and water saturation.

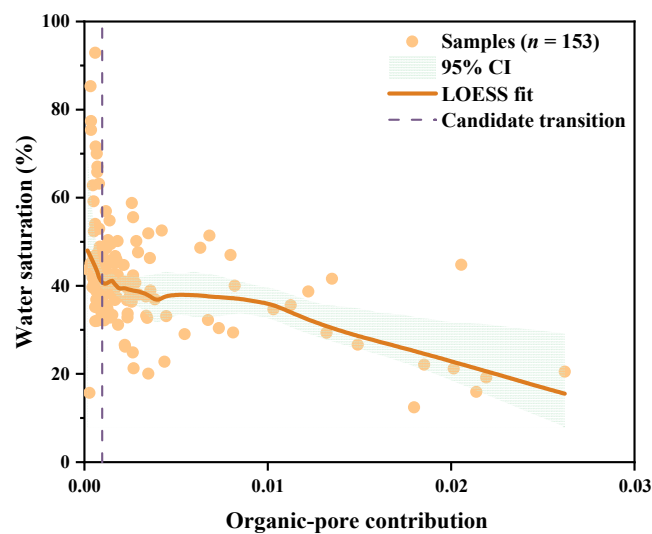


Figure 9. Relationship between organic-pore contribution ratio and water saturation.

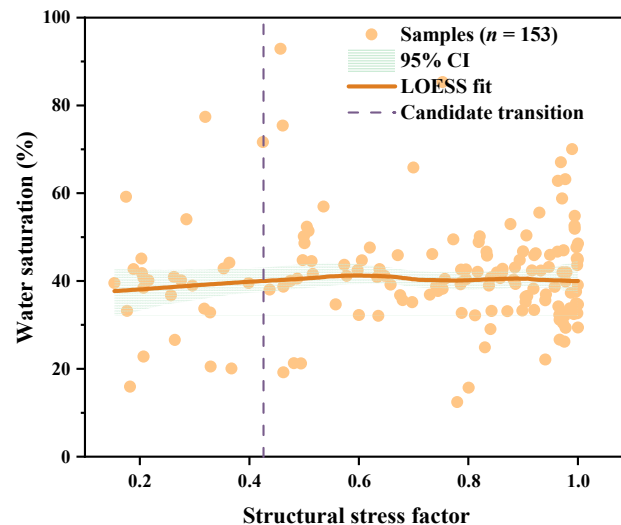


Figure 10. Relationship between structural-stress factor and water saturation.

3.2.1. Clay Content

Clay content exhibited a moderate positive association with measured water saturation (Spearman's $\rho = 0.393$, $p < 0.001$; Figure 6). The LOESS curve increased across most of the observed clay-content range and gradually leveled off within the clay-rich interval. Piecewise regression identified a candidate transition at approximately 54%. However, its bootstrap 95% confidence interval was relatively wide (30–56%), the improvement over the linear model was limited ($\Delta R^2 = 0.011$), and only 26 samples occurred above the estimated transition. The previously proposed value of approximately 35% should therefore not be interpreted as a critical clay-content threshold. The results instead support a broad, dataset-dependent transition within the clay-rich interval.

This relationship is geologically plausible because clay-rich shale generally contains a larger specific surface area and a higher proportion of clay-bound water. Surface-charge effects, capillary confinement, and clay interlayer water may consequently increase the apparent water saturation response. Nevertheless, the gradual plateau at high clay contents indicates that the effect of clay is also modified by pore structure, compaction, mineral composition, and associated changes in pore connectivity. The statistical relationship therefore supports an association between clay enrichment and water saturation but does not demonstrate that clay content alone directly determines water saturation.

3.2.2. Structural Curvature

The water saturation distributions differed between the two structural-curvature zones (Figure 7). Samples in the low-curvature zone had a mean water saturation of 43.41% and a median of 41.50% ($n = 102$), whereas the high-curvature zone had a lower mean of 37.51% and a median of 36.80% ($n = 51$). The difference was statistically detectable using the Mann–Whitney test ($p = 0.0058$). In addition, water saturation showed greater dispersion in the high-curvature zone, indicating increased reservoir heterogeneity.

The observed distribution is consistent with the possibility that deformation-related heterogeneity modifies fracture development, pore-network connectivity, and fluid redistribution. However, structural curvature in the present analysis was represented by a log-derived proxy rather than direct seismic or structural measurements. Consequently, the result should be interpreted as an association between structural heterogeneity and water saturation. It does not demonstrate that increasing structural curvature directly increases or decreases reservoir water content, nor does it independently establish a causal curvature–clay interaction.

3.2.3. Organic-Matter Abundance (TOC)

TOC was negatively associated with measured water saturation (Spearman's $\rho = -0.312$, $p < 0.001$; Figure 8). The LOESS curve displayed the greatest variation in the low-TOC interval and became more gradual at higher TOC values. Piecewise regression identified a candidate transition at approximately 0.87%, with a bootstrap 95% confidence interval of 0.79–2.68%. The piecewise model improved R^2 by 0.053 relative to the linear model.

These results do not support the previously proposed TOC threshold of approximately 5%. Instead, the available samples indicate a gradual and nonlinear association concentrated mainly within the lower-TOC range. TOC primarily represents the bulk abundance of organic matter rather than the geometry or connectivity of the organic-pore system. Its relationship with water saturation may reflect the combined influence of organic-matter abundance, thermal evolution, wettability, and hydrocarbon occupancy. Because these effects cannot be separated using the present univariate relationship, the candidate transition should be regarded as exploratory and dataset-dependent rather than as a universal geological threshold.

3.2.4. Organic-Pore Contribution Ratio

The organic-pore contribution descriptor showed a significant negative association with measured water saturation (Spearman's $\rho = -0.401$, $p < 0.001$; Figure 9). Piecewise regression identified a candidate transition at approximately 9.73×10^{-4} , with a bootstrap 95% confidence interval of 5.90×10^{-4} – 2.20×10^{-3} . The resulting R^2 improvement of 0.081 was the largest among the examined candidate transitions, indicating a comparatively clear nonlinear change within the current dataset.

Unlike TOC, which primarily characterizes organic-matter abundance, the organic-pore contribution descriptor represents the participation of organic matter in the effective pore system. Because the descriptor integrates TOC, porosity, and clay content, it provides information on pore-space allocation rather than organic-matter quantity alone. Samples with similar TOC values may therefore exhibit different organic-pore contributions because of differences in total porosity and clay-related pore modification. The observed relationship with water saturation suggests that the model captures the combined influence of organic-associated pore space and water-accessible pore structure. However, because this descriptor is derived rather than independently measured, its contribution should not be interpreted as direct causal evidence that organic pores reduce water saturation.

3.2.5. Structural-Stress Factor

In contrast to clay content, TOC, and the organic-pore contribution descriptor, the structural-stress factor showed no significant monotonic relationship with measured water saturation (Spearman's $\rho = 0.018$, $p = 0.821$; Figure 10). Although piecewise regression produced a numerical breakpoint at approximately 0.426, its bootstrap 95% confidence interval extended from 0.323 to 0.975, and the improvement over the linear model was only $\Delta R^2 = 0.016$.

The current data therefore do not support a stable optimum within the previously proposed range of 0.6–0.8 or a systematic reversal above 0.9. The importance of the structural-stress factor in the multivariable model is more appropriately attributed to its nonlinear interactions with clay content, pore properties, depth, and structural heterogeneity. Accordingly, the candidate breakpoint shown in Figure 10 is retained only as a descriptive statistical result and is not assigned a direct geomechanical or causal interpretation.

3.3. Engineering Implementation and Field Deployment Considerations

The computational performance was evaluated on a workstation equipped with a 13th Gen Intel Core i7-13700K processor and 64 GB RAM without GPU acceleration. Raw-data integration and preprocessing for the 153 samples required 12.97 s, whereas complete five-fold model training required 3.58 s after construction of the integrated sample matrix. The

complete offline workflow therefore required approximately 16.55 s. End-to-end inference required 47.7 ms for a single sample and 68.5 ms for a batch of 1000 depth points, indicating that model inference is unlikely to constitute the principal computational limitation in field applications.

For field implementation, depth-indexed data can be imported in LAS, WITSML, or CSV format and subjected to automated checks for units, depth alignment, missing values, and physically implausible records. The validated inputs are then processed using the fixed preprocessing and feature-generation parameters before ensemble prediction. Because conventional logging variables can be obtained more rapidly than laboratory-derived properties, a two-stage workflow is recommended: an initial prediction using available logging and well-level information, followed by an updated prediction when interpreted petrophysical or laboratory data become available. Predictions associated with incomplete or out-of-range inputs should be flagged for manual review.

4. Conclusions

This study developed a feature-level physics-guided machine-learning framework, PhysML-Hybrid, for water saturation prediction in deep Silurian shale gas reservoirs. The principal conclusions are as follows:

1. Five groups of physics-derived descriptors were constructed to represent clay–water interfacial behavior, low-resistivity effects, pore connectivity, organic-pore development, and structural stress. Their integration with conventional reservoir variables enabled relevant geological and petrophysical information to be incorporated into the input feature space. The validation-weighted RF–XGB–BNN ensemble further combined the complementary predictive capabilities of heterogeneous learners.
2. Target-stratified five-fold cross-validation conducted exclusively within the 107-sample training set yielded mean R^2 , MAE, and RMSE values of 0.902 ± 0.010 , $1.77\% \pm 0.12\%$, and $2.34\% \pm 0.12\%$, respectively. On the fixed 30-sample independent test set, the final model achieved an R^2 of 0.902, an MAE of 1.77%, and an RMSE of 2.34%. The close agreement between the validation and independent test results indicates limited performance deterioration for samples excluded from model development, although the relatively small test set requires a cautious interpretation of generalization.
3. Comparisons with the conventional Archie model, SVM, ML-only ensemble, and Phy-XGB showed that PhysML-Hybrid provided the best overall predictive performance. The improvements over ML-only and Phy-XGB indicate that both the physics-derived descriptors and the heterogeneous ensemble strategy contributed to the final performance. The three blind-well cases produced relative errors of 3.09–9.97%, providing preliminary supplementary evidence of cross-well applicability rather than definitive cross-block validation.
4. SHAP and complementary statistical analyses identified clay content as the dominant predictor, followed by the structural-curvature–clay interaction, TOC, pore connectivity, and structural descriptors. These relationships are physically consistent with the combined effects of clay-bound water, organic-matter development, pore-network properties, and structural deformation. Nevertheless, the estimated breakpoints represent dataset-specific candidate transitions, and the SHAP results should not be interpreted as direct evidence of geological causality.
5. The CPU-based implementation provided sufficiently rapid inference for batch processing and potential well-site application. However, the study remains limited by its relatively small dataset, restricted geographical coverage, and dependence on the availability and quality of structural descriptors. Moreover, although the BNN component used the posterior mean as its point prediction, prediction-interval coverage and calibration were not quantitatively evaluated. The current results should therefore be

interpreted as deterministic prediction-performance evidence rather than a calibrated probabilistic reliability assessment. Future work should include larger independent multi-well and multi-basin datasets, external validation, local recalibration, predictive-uncertainty analysis, and reduced-input configurations for field conditions with incomplete measurements.

Author Contributions: Conceptualization, G.Z. and L.X.; methodology, G.Z., L.X. and H.C.; software, G.Z. and Y.D.; validation, G.Z., H.C. and R.W.; formal analysis, G.Z. and D.T.; investigation, G.Z., H.C. and R.W.; resources, L.X. and M.W.; data curation, G.Z. and Y.D.; writing—original draft preparation, G.Z.; writing—review and editing, L.X., H.C., R.W., Y.D., D.T. and M.W.; visualization, G.Z. and D.T.; supervision, L.X.; project administration, L.X. and M.W.; funding acquisition, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of the People’s Republic of China, National Key Science and Technology Project for New Oil and Gas Exploration and Development, grant number 2025ZD1404100; the National Natural Science Foundation of China, grant number 52274048; the Ministry of Science and Technology of the People’s Republic of China the National Science and Technology Major Project of China, grant number 2025ZD1405202; and the Beijing Capital High-End Leading Talent Aggregation and Cultivation Project, grant number 202504841068.

Data Availability Statement: The data presented in this study are available on request from the corresponding author (The data presented in this study are available from the corresponding author upon reasonable request. The data are not publicly available because they contain confidential field information provided by the collaborating company and are subject to data-sharing restrictions.).

Conflicts of Interest: Author Gaofeng Zou was employed by the Exploration and Development Research Institute of Jiangnan Oilfield Branch, Sinopec, and author Ruyue Wang was employed by the Research Institute of Petroleum Exploration and Development, Sinopec Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wang, Y.; Wei, G.; Shen, J.; Qiu, Z.; Li, X.; Zhang, Q.; Zhang, L.; Wang, C.; Liu, W. Analysis on Carbonization Distribution and Main Controlling Factors of Organic Matter in Marine Shale in Sichuan Basin and Its Periphery. *Nat. Gas Geoscience* **2022**, *33*, 843–859.
2. Li, J.Q.; Lu, S.F.; Zhang, P.F.; Li, W.B.; Jing, T.Y.; Feng, W.J. Quantitative Characterization and Microscopic Occurrence Mechanism of Pore Water in Shale Matrix. *Acta Pet. Sin.* **2020**, *41*, 979. <https://doi.org/10.7623/syxb202008007>.
3. Wei, F.; Liu, Z.; Chen, F.; Yan, W.; Wang, Q. Discussion on Genesis and Geological Significance of “Low Resistivity and Low Gas Content” of Longmaxi Formation Shale in Southeastern Sichuan. *Exp. Pet. Geol.* **2023**, *45*, 1089–1096. <https://doi.org/10.11781/sysydz2023061089>.
4. Liu, X.; Qi, L.; Nie, B.; Li, C.; Du, F.; Zhang, L.; Zhao, X.; Wei, S.; Li, Z. Effects of Al₂O₃ Nanofluids and pH Conditions on Wettability and Micromechanical Properties of Coal. *J. Rock Mech. Geotech. Eng.* **2026**, *in press*. <https://doi.org/10.1016/j.jrmge.2026.03.047>.
5. Qi, L.; Liu, X.; Nie, B.; Li, J.; Chen, P.; Han, H.; Li, Z.; Wei, S.; Deng, B. Mechanisms of Enhanced Wettability and Nanomechanical Strength in Soft Coal Seams Modified by Acidic SiO₂ Nanofluids. *Int. J. Min. Sci. Technol.* **2026**, *36*, 849–872. <https://doi.org/10.1016/j.ijmst.2026.03.001>.
6. Liu, X.; Jia, X.; Yang, T.; Nie, B.; He, C.; Li, C.; Bao, S. Permeability Enhancement Mechanism in Coal Beds Fractured Using High-Voltage Electrical Pulse Treatment: Cross-Scale Transportation of Coalbed Methane. *Green Smart Min. Eng.* **2025**, *2*, 402–412. <https://doi.org/10.1016/j.gsme.2025.10.004>.
7. Xu, S.Y.; Zeng, Y.Y.; Lin, Y.; Zhu, Y.; Xiao, X.W.; Li, T.J.; Shan, S.J.; Ma, Z.K. Chemical Characteristics and Genesis and Evolution of Formation Water in the Middle Permian Maokou Formation, Central Sichuan Basin. *China Pet. Explor.* **2024**, *29*, 116–129. <https://doi.org/10.3969/j.issn.1672-7703.2024.06.009>.
8. Tang, X.L.; Jiang, Z.X.; Shi, Y.G.; Wang, Y.N. Gas-Water Storage State and Genetic Mechanism of Shale Gas. In Proceedings of the 33rd National Natural Gas Academic Annual Conference (2023), Nanning, China, 1 June 2023.

9. Archie, G.E. The Electrical Resistivity Log as an Aid in Determining Some Reservoir Characteristics. *Trans. AIME* **1942**, *146*, 54–62. <https://doi.org/10.2118/942054-G>.
10. Waxman, M.H.; Smits, L.J.M. Electrical Conductivities in Oil-Bearing Shaly Sands. *Soc. Pet. Eng. J.* **1968**, *8*, 107–122. <https://doi.org/10.2118/1863-A>.
11. Clavier, C.; Coates, G.; Dumanoir, J. Theoretical and Experimental Bases for the Dual-Water Model for Interpretation of Shaly Sands. *Soc. Pet. Eng. J.* **1984**, *24*, 153–168. <https://doi.org/10.2118/6859-PA>.
12. Iqbal, M.A.; Rezaee, R. Porosity and Water Saturation Estimation for Shale Reservoirs: An Example from Goldwyer Formation Shale, Canning Basin, Western Australia. *Energies* **2020**, *13*, 6294. <https://doi.org/10.3390/en13236294>.
13. Deng, Y.; Guo, R.; Tian, Z.; Zhao, L.; Hu, D.; Liu, H.; Liu, Y. Water Saturation Modeling Using Modified J-Function Constrained by Rock Typing Method in Bioclastic Limestone. *Oil Gas Sci. Technol.–Rev. IFP Energ. Nouv.* **2020**, *75*, 66. <https://doi.org/10.2516/ogst/2020059>.
14. Brownstein, K.R.; Tarr, C.E. Importance of Classical Diffusion in NMR Studies of Water in Biological Cells. *Phys. Rev. A* **1979**, *19*, 2446–2453. <https://doi.org/10.1103/PhysRevA.19.2446>.
15. Coates, G.R.; Xiao, L.; Prammer, M.G. *NMR Logging: Principles and Applications*; Halliburton Energy Services: Houston, TX, USA, 1999. ISBN 978-0-9679026-0-9.
16. Kenyon, W.E. Nuclear Magnetic Resonance as a Petrophysical Measurement. *Nucl. Geophys. (Int. J. Radiat. Appl. Instrum. Part E)* **1992**, *6*, 153–171.
17. Sun, J.; Yao, Y.; Sun, F.; Su, J.; Lu, J.; Liu, K.; Chi, P. NMR Evaluation of Shale Oil Mobility: Combined Pyrolysis and CO₂ Huff-N-Puff. *Appl. Sci.* **2024**, *14*, 11251. <https://doi.org/10.3390/app142311251>.
18. Wang, X.; Xie, R.; Guo, J.; Tian, B. A Novel Intelligent Evaluation Method for Fluid Saturation in Shale Oil Reservoirs Using NMR T1-T2 Spectra. *Fuel* **2026**, *405*, 136809. <https://doi.org/10.1016/j.fuel.2025.136809>.
19. Zhang, Z.; Song, Y.; Yin, S.; Liu, B.; Zhao, H.; Yan, W.; Zheng, J. Analysis of the Controlling Factors for Sweet Spots in Lacustrine Shale Based on Core Experiments and 2D NMR Logging. *Sci. Rep.* **2025**, *15*, 20440. <https://doi.org/10.1038/s41598-025-04771-1>.
20. Li, Y.; Hu, Z.; Cai, C.; Liu, X.; Duan, X.; Chang, J.; Li, Y.; Mu, Y.; Zhang, Q.; Zeng, S.; et al. Evaluation Method of Water Saturation in Shale: A Comprehensive Review. *Mar. Pet. Geol.* **2021**, *128*, 105017. <https://doi.org/10.1016/j.marpetgeo.2021.105017>.
21. Zhao, H.; Zheng, J.; Zhang, Z.; Qin, H.; Wang, C.; Zhu, J.; Chen, A.; Chen, L.; Liu, P.; Yang, W.; et al. Research Progress and Prospects of Oil Saturation Evaluation Methods in Shale Oil Reservoirs. *Processes* **2024**, *12*, 2421. <https://doi.org/10.3390/pr12112421>.
22. Singh, H.; Seol, Y.; Myshakin, E.M. Prediction of Gas Hydrate Saturation Using Machine Learning and Optimal Set of Well-Logs. *Comput. Geosci.* **2021**, *25*, 267–283. <https://doi.org/10.1007/s10596-020-10004-3>.
23. Neff, P.; Steineder, D.; Stummer, B.; Clemens, T. Estimation of Initial Hydrocarbon Saturation Applying Machine Learning Under Petrophysical Uncertainty. *SPE Reserv. Eval. Eng.* **2021**, *24*, 325–340. <https://doi.org/10.2118/203384-PA>.
24. Kumar, A.; Rodrigues, P.; Kareem, A.K.; Sekac, T.; Abdullaev, S.; Chohan, J.S.; Manjunatha, R.; Rethik, K.; Dasi, S.; Kiani, M. Modeling Saturation Exponent of Underground Hydrocarbon Reservoirs Using Robust Machine Learning Methods. *Sci. Rep.* **2025**, *15*, 373. <https://doi.org/10.1038/s41598-024-84556-0>.
25. Ibrahim, A.F.; Elkhatny, S.; Abdelraouf, Y.; Al Ramadan, M. Application of Various Machine Learning Techniques in Predicting Water Saturation in Tight Gas Sandstone Formation. *J. Energy Resour. Technol.* **2022**, *144*, 083009. <https://doi.org/10.1115/1.4053248>.
26. Mellal, I.; Latrach, A.; Rasouli, V.; Bakelli, O.; Dehdouh, A.; Ouadi, H. Water Saturation Prediction in the Middle Bakken Formation Using Machine Learning. *Eng* **2023**, *4*, 1951–1964. <https://doi.org/10.3390/eng4030110>.
27. Akbari, A.; Ranjbar, A.; Kazemzadeh, Y.; Martyushev, D.A. Enhanced Water Saturation Estimation in Hydrocarbon Reservoirs Using Machine Learning. *Sci. Rep.* **2025**, *15*, 29846. <https://doi.org/10.1038/s41598-025-13982-5>.
28. Wan, C.-X.; Guo, X.-S.; Shen, B.-J.; Song, Y.; Wang, P.-W.; Wang, R.-Y. DLVO-Based Estimates of Critical Water Saturation of Shale: A Case Study of the Lower Silurian Longmaxi Shale in Sichuan Basin. *Pet. Sci.* **2025**, *22*, 1876–1888. <https://doi.org/10.1016/j.petsci.2025.03.003>.
29. Han, J.-X.; Xue, L.; Wei, Y.-S.; Qi, Y.-D.; Wang, J.-L.; Liu, Y.-T.; Zhang, Y.-Q. Physics-Informed Neural Network-Based Petroleum Reservoir Simulation with Sparse Data Using Domain Decomposition. *Pet. Sci.* **2023**, *20*, 3450–3460.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.