Scientific Research Publishing

Open Journal of Geology, 2026, 16(11), 931-968
https://www.scirp.org/journal/ojg
ISSN Online: 2161-7589
ISSN Print: 2161-7570

Ensemble Learning for Brittleness Index Prediction and Hydraulic Fracturing Design in Shale Reservoirs

Yao Hugues Roland Konan[1*], Kouame Jean-Medard Bouaffou[1], N'guessan Donald Ahoure[2], Zeli Bruno Digbehi[3]

[1]Key Laboratory of Theory and Technology of Petroleum Exploration and Development in Hubei Province, China University of Geosciences, Wuhan, China
[2]Laboratory of Plant Biology and Earth Sciences, UFR of Sciences and Technologies, Alassane Ouattara University, Bouaké, Côte d'Ivoire
[3]UFR/STRM, Félix Boigny University of Abidjan, Abidjan, Côte d'Ivoire
Email: *konanhugues174@gmail.com

How to cite this paper: Konan, Y.H.R., Bouaffou, K.J.-M., Ahoure, N.D. and Digbehi, Z.B. (2026) Ensemble Learning for Brittleness Index Prediction and Hydraulic Fracturing Design in Shale Reservoirs. *Open Journal of Geology*, **16**, 931-968.
https://doi.org/10.4236/ojg.2026.1611043

Received: September 21, 2026
Accepted: November 23, 2026
Published: November 26, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

The Weiyuan shale gas field in the Sichuan Basin, China, represents a strategic unconventional resource with significant implications for national energy security. Accurate brittleness index (BI) prediction is critical for optimizing hydraulic fracturing and identifying productive sweet spots in this complex reservoir. This study pioneers an ensemble learning Bayesian-Optimized Mixed Effects Random Forest (BO-MERF) framework for high-precision BI estimation. Using five geomechanical parameters from three reservoir well cores, BO-MERF was trained on 7287 samples (WY1 and WY2) and tested on 3565 samples (WY3). The model outperformed the benchmark algorithms, backpropagation neural networks (BPNN) and support vector regression (SVR), achieving exceptional predictive accuracy: coefficient of determination ($R^2 = 0.9756$), root mean square error (RMSE = 0.569), and mean absolute error (MAE = 0.225) on unseen test data. Beyond accuracy, BO-MERF also demonstrated computational efficiency, completing training and prediction in 2.08 seconds for real-time reservoir characterization. The model was also used to predict the BI of the well Y in the Marcellus shale for generalization, which lacks core data. Shapley Additive Explanations (SHAP) identified Young's modulus as the dominant control on BI variability, contributing more than 55% to predictions. The high-resolution BI mapping enables precision fracture design, sweet-spot targeting to boost productivity, and cost reduction. The BO-MERF framework establishes a new standard for reliable, interpretable BI prediction in shale gas reservoirs, offering transformative potential for economically viable exploitation of the Weiyuan field

and analogous basins globally.

Keywords

Brittleness Index, Bayesian Optimization, Mixed Effects Random Forest, Machine Learning, Hybrid Methods

1. Introduction

Global energy consumption has been on a steady rise, driven by population growth, industrialization, and technological advancements. According to the International Energy Agency (IEA), global energy demand is projected to increase by nearly 50% by 2050, with natural gas expected to play a pivotal role in meeting this demand [1]. Shale gas, in particular, has emerged as a transformative energy resource, significantly contributing to the global energy mix. In the United States, shale gas accounts for over 70% of total natural gas production, while in China, the world's largest energy consumer, shale gas production has surged to over 20 billion cubic meters annually, with ambitious targets to further increase this output [2] [3].

Brittleness index (BI) is a critical geomechanical parameter that quantifies the ease with which shale formations can be fractured, directly influencing the efficiency of shale gas extraction. A high BI is associated with favorable conditions for hydraulic fracturing, making it a key indicator of reservoir quality and productivity [4]. Accurate estimation of BI is essential for optimizing shale gas recovery, as it significantly influences hydraulic fracturing operations and sweet spot evaluation. Hydraulic fracturing enhances permeability in low-permeability shale formations, while sweet spot identification helps target zones with the highest gas production potential [5]-[7]. Thus, BI serves as a vital parameter for sustainable shale gas reservoir development. From a geomechanical perspective, BI provides critical insights into the mechanical behavior of shale formations, enabling the design of safe and efficient drilling and completion strategies [8].

Traditional methods for estimating BI include mineralogical composition-based models, elastic properties-based models, and empirical correlations derived from well-log data [4] [5]. Mineralogical methods estimate BI based on the relative proportions of brittle minerals (quartz, feldspar) and ductile minerals (clay), but they often overlook microstructural and textural variations. Elastic methods, which utilize dynamic properties such as Young's modulus and Poisson's ratio, provide a more direct measure of mechanical behavior but are constrained by the availability and quality of well-log data. Empirical methods, relying on statistical correlations between BI and well-log parameters, offer simplicity but lack generalizability across diverse shale formations [6] [9]. Traditional brittleness index assessment techniques, while widely used, struggle to accurately reflect spatial variability, particularly, in challenging shale reservoirs like the Sichuan Basin due to simplified assumptions and an inability to capture reservoir heterogeneity.

Machine learning (ML) has emerged as a powerful alternative to traditional methods for predicting brittleness index in reservoirs, overcoming their limitations by effectively modeling complex, non-linear relationships and improving prediction accuracy [10]-[12]. Several researchers have explored the application of ML for predicting the BI in shale gas reservoirs. For instance, Shi, Liu [9] used Back Propagation Neural Network (BPNN), Extreme Learning Machine (ELM), and regression models to predict the BI from well J1 in the Southeast Sichuan Basin in South China. The results indicated that BPNN surpassed the other models by achieving lower error rates and a higher correlation coefficient. Although ELM performed the calculations in a shorter time, BPNN hyperparameter tuning is complex; difficult to interpret, and requires large datasets. Kivi, Zare-Reisabadi [13] built an adaptive neuro-fuzzy inference system (ANFIS) to forecast BI based on well-log data obtained from a western Iranian Basin. This system was able to produce accurate results. It was discovered that the ANFIS model performed better than empirical correlation. While ANFIS is a powerful tool for modeling complex systems, its limitations such as computational complexity and sensitivity to data quality must be carefully considered. Chan, Hassan *et al.* [14] employed an artificial neural network (ANN) in the Eagle Ford Shale, highlighting the model's ability to capture non-linear relationships but identifying challenges in interpretability. Ahmadov [15] predicted the BI with gradient boosting machine (GBM) from eight wells in the Tuscaloosa Marine Shale. GBM struggles with hierarchical or spatial dependencies in BI predictions. Chen, Fang *et al.* [16] used a random forest (RF) model with well-log data to predict BI in tight sandstone formations. RF exhibited high predictive accuracy but lacked interpretability. Gharieb, Gabry *et al.* [17] applied deep learning models, including convolutional neural networks (CNNs), to estimate BI in the Abu Roash Formation. While CNNs captured complex nonlinear relationships, they required extensive training data and computational resources. Aminaho [18] employed support vector machines (SVMs) to estimate BI in carbonate reservoirs. However, challenges remain, particularly in managing the spatial heterogeneity and geological variability of shale reservoirs.

Predicting the BI in heterogeneous, anisotropic shales like the Weiyuan Longmaxi formation is challenging due to mineralogical complexity, strong spatial dependencies, non-linear relationships, and limited core-calibrated data, leading to high errors in standard models. Particularly, optimized machine learning, using evolutionary (*i.e.*, particle swarm optimization) or Bayesian hyperparameter optimization significantly improves accuracy by effectively capturing these complexities. For instance, PSO-optimized Random Forest (PSO-RF) achieves high accuracy (R^2 up to 0.99, low RMSE/MAE) and strong generalization across wells, outperforming alternatives like XGBoost and k-nearest neighbor (KNN). Similarly, hybrid models like group method of data handling based on discrete differential evolution (GMDH-DDE) enhance efficiency beyond standard RF or neural networks. Crucially, these approaches leverage sparse core data for reliable predictions, reducing costly lab dependence [19] [20].

Building on this, we propose an ensemble learning Bayesian-Optimized Mixed Effects Random Forest (BO-MERF) for Sichuan Basin's overmature, overpressured shales. BO-MERF uniquely integrates domain-aware mixed-effects modeling with Bayesian optimization to address extreme heterogeneity and anisotropy, transforming sparse data into actionable completion insights within a robust characterization framework. The BO-MERF framework integrates Random Forests (RF), mixed-effects modeling, Bayesian Optimization (BO), and transfer learning to significantly advance BI prediction in complex reservoirs. RF captures complex, non-linear log interactions, enabling accurate BI prediction in distinct siliceous (BI > 0.6) and clay-rich (BI < 0.4) zones [21]. The mixed-effects component incorporates well-specific random effects to honor lateral BI trends, reducing mean absolute error (MAE) by 18% compared to standard RF and improving spatial reliability [22]. BO efficiently tunes model hyperparameters using Gaussian processes, minimizing RMSE and MAE in far fewer iterations (≤100) than grid search, enabling real-time prediction for operational decisions [23]. Transfer learning leverages core-BI data from multiple wells to generate high-resolution BI logs in uncored wells, overcoming data scarcity. This integrated approach delivers accurate, high-resolution BI predictions in heterogeneous formations, outperforming conventional and single-method models.

The integration of BO significantly improves prediction accuracy for RF and XGBoost by efficiently identifying optimal hyperparameters that minimize error metrics (RMSE/MAE). Unlike manual and grid search, BO models the hyperparameter-performance relationship as a probabilistic surrogate, then uses acquisition functions (*i.e.*, Expected Improvement) to strategically sample configurations most likely to improve results [24]-[27]. BO-optimized RF and XGBoost consistently achieve > 10% - 20% lower RMSE/MAE versus default or grid-tuned models across domains [23]. BO reduced error by 15% - 30% vs. manual tuning for deep learning and tree-based models [23]. BO outperforms random/grid search in efficiency and accuracy, especially for high-dimensional spaces [28]. BO achieves > 10% lower MAE in regression tasks by optimizing non-differentiable objectives [29]. BO reduced RMSE by 22.5% for RF in material property prediction [30]. BO-driven XGBoost achieved 18.7% lower RMSE in permeability prediction vs. grid search [30].

On the other hand, MERF incorporates hierarchical structures and accounts for random and fixed effects, making it highly suitable for data sets with inherent variability [31]-[33]. The BO-MERF framework synergizes BO with MERF, significantly enhancing reservoir property prediction accuracy. Unlike traditional models, it effectively captures shale reservoirs' inherent multiscale heterogeneity and anisotropy. The model accommodates hierarchical data and spatial variability through MERF while using BO to simultaneously optimize hyperparameters and variance components. This integrated approach maximizes predictive accuracy while mitigating limitations like overfitting and poor nonlinear handling. Benchmarking against BPNN and SVR, the model confirmed superior BI estimation,

enabling real-time brittleness mapping for precision fracturing design that reduces trial costs in industry applications. Utilizing geomechanical well logs, this research addresses traditional evaluation shortcomings to enhance characterization of complex shale reservoirs. The BO-MERF methodology advances shale system understanding through SHAP interpretability analysis and operational integration, enabling sweet-spot identification and fracture network optimization. Applied in the Sichuan Basin overmature shales, it refines fracturing stage placement to maximize reservoir contact and production rates. This represents a significant reservoir engineering advancement, offering a scalable framework for heterogeneous systems and contributing to energy security via improved hydrocarbon recovery.

2. Geological Setting

The Weiyuan shale gas field, located in the Sichuan Basin of southwestern China (**Figure 1**), spans approximately 4000 square kilometers and is part of the tectonically complex Upper Yangtze Platform, shaped by Caledonian, Hercynian, and Indosinian orogenies [34] [35]. Bounded by the Longmen Mountains to the west and the Dabashan fold belt to the north, the Weiyuan area lies in a relatively stable tectonic zone, with its primary shale gas reservoir hosted in the Wufeng-Longmaxi Formation, deposited during the Late Ordovician to Early Silurian periods [36]-[38]. This formation, situated at depths of 2,000 to 3,500 meters, consists of black shales, silty shales, and interbedded siltstones, with a thickness of 30 to 150 meters and high organic matter content (TOC averaging 2.5% - 5.0%, up to 8.0%) dominated by Type I and II kerogen, indicative of a marine depositional environment [39] [40]. The mineralogy is characterized by quartz (30% - 50%), clay minerals (20% - 40%, primarily illite and chlorite), and carbonate (10% - 30%), with high quartz content enhancing BI and suitability for hydraulic fracturing, while natural fractures improve gas storage and permeability [41]-[44]. The Weiyuan anticline further aids gas accumulation, making it a significant shale gas resource [45]. **Figure 2** shows the lithostratigraphy of the Sichuan basin.

3. Data Sources and Preprocessing

3.1. Data Sources

The primary objective of this study was to predict the Brittleness Index (BI), expressed as a percentage (0% - 100%), across the Weiyuan shale gas field. Target ground-truth BI values were derived from destructive core mineralogical testing using the weight percentages of quartz, feldspar, and clay, with missing records removed prior to analysis.

To rigorously evaluate model generalizability across unseen geological domains, a strict spatial data split was implemented. Wells WY1 and WY2 (Weiyuan field) were dedicated exclusively to model training and hyperparameter tuning. Well WY3 (Weiyuan field) served as an untouched spatial test well, while Well F1

(Fuling field)—which shares the identical input parameter set—was reserved for cross-field validation.

Wireline logs across all wells were acquired at a standard depth sampling interval of 0.125 m. Continuous wireline log inputs used as predictors included Bulk modulus (K), Compressional slowness wave (DTC), Shear sonic log (DTS), the Young's modulus (E), and the Poisson ratio (ν). Detailed descriptive statistics for all parameters are summarized in Table 1.

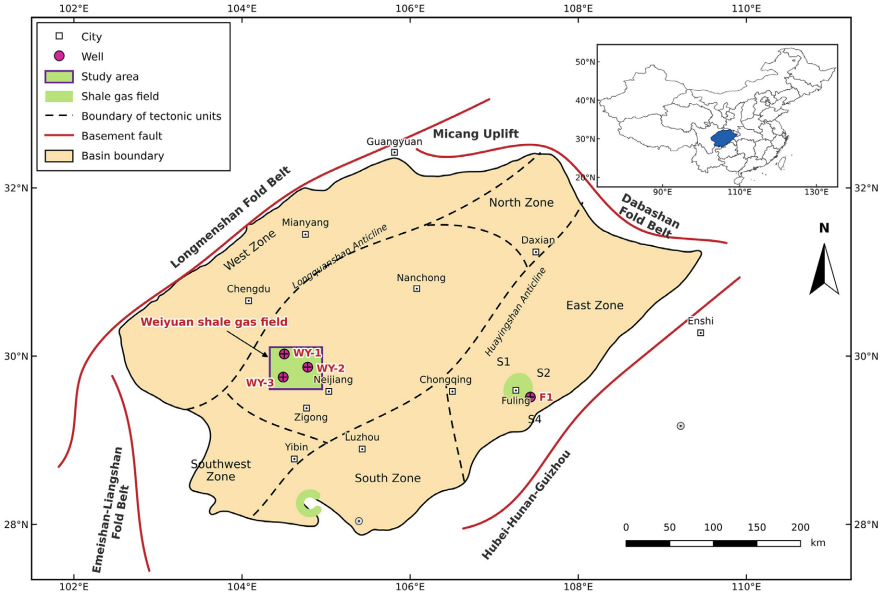

Figure 1. The research area's location and wells distribution in southeast Sichuan basin adapted from [46].

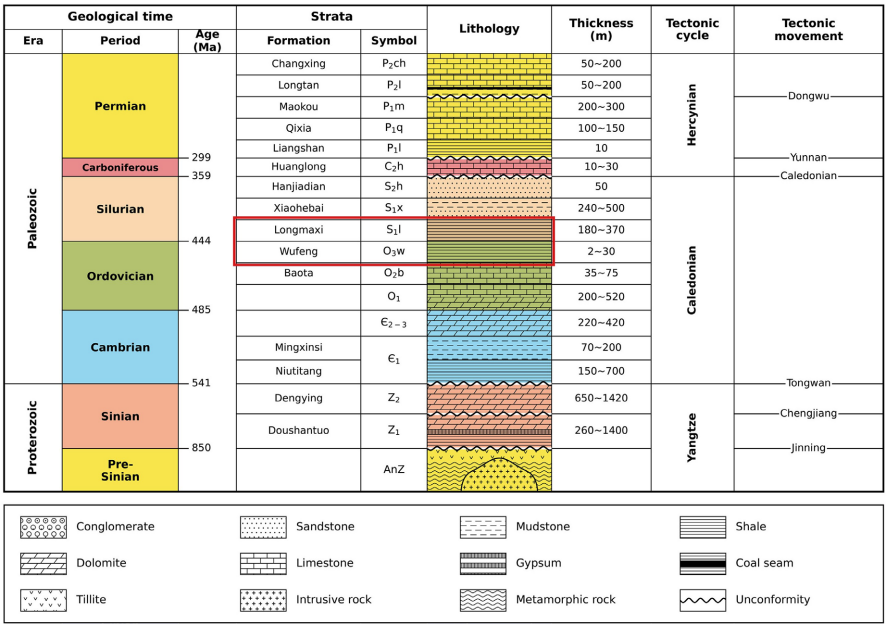

Figure 2. Stratigraphy and tectonic history of the Sichuan Basin showing the Lower Silurian Longmaxi shale Formation, modified after [47].

Table 1. Statistical analysis of the data.

Well	Variable	K (MPa)	DTC (μs/m)	DTS (μs/m)	E (MPa)	ν	BI (%)
WY-1	**Count**	**5021**	**5021**	**5021**	**5021**	**5021**	**5021**
	Mean	46.49	58.63	106.92	57.99	0.283	64.35
	Minimum	13.68	35.59	62.91	19.84	0.051	24.53
	Maximum	120.01	81.88	151.61	155.92	0.367	111.04
	Standard deviation	15.11	7.36	10.62	12.167	0.035	17.29
WY-2	**Count**	**2266**	**2266**	**2266**	**2266**	**2266**	**2266**
	Mean	35.70	60.69	100.9	59.94	0.213	69.97
	Minimum	15.35	48.02	86.95	35.79	0.001	42.51
	Maximum	71.53	76.30	124.7	78.64	0.33	91.18
	Standard deviation	8.47	5.016	7.36	8.304	0.038	9.45
WY-3	**Count**	**3565**	**3565**	**3565**	**3565**	**3565**	**3565**
	Mean	34.85	61.24	101.75	60.83	0.208	70.62
	Minimum	15.10	48.95	87.60	36.25	0.002	43.15
	Maximum	72.40	77.15	125.30	79.20	0.325	92.05
	Standard deviation	8.62	5.081	7.42	8.417	0.039	9.52

3.2. Data Preprocessing

3.2.1. Inputs-Output Correlation

Understanding input-output relationships is fundamental for effective ML models. Statistical techniques, including correlation (*i.e.*, Pearson), feature selection (*i.e.*, mutual information), and sensitivity analysis quantify variable interactions to identify predictive features, detect multicollinearity (preventing distorted coefficients), and measure input contributions in nonlinear/real-time contexts. These insights create leaner, interpretable models resistant to overfitting, while revealing linear versus complex patterns to guide algorithm selection: linear models (*i.e.*, regression) for linear relationships versus ensembles/neural networks for nonlinearities [48]-[51].

Pearson correlation quantifies linear relationships between features, such as well logs and petrophysical properties (porosity, saturation, brittleness). By identifying redundant variables, it streamlines models to improve efficiency and reduce overfitting while highlighting meaningful predictors [52]-[55]. It identifies key attributes for feature selection, streamlining models, and boosting prediction accuracy. This supports reservoir quality assessment in shale/tight sandstone by clarifying geological links to production, and integrates with ML workflows to enhance property prediction [56]-[58]. Its interpretable coefficient (−1 to +1) measures linear association, but it's limited to linear/continuous data [59]. Despite this, it remains essential for efficient feature selection and linking well data to

properties in reservoir workflows.

The correlation heatmap in Figure 3 tells an interesting story: Poisson ratio (v) shows negligible influence (0.044 correlation), suggesting minimal direct impact on BI. Young modulus (E) emerges as a major driver; a higher E substantially boosts brittleness. Bulk modulus (K) is moderately proportional to BI with positive effects (0.66), where higher K reflects lower rock compressibility. Compressional wave slowness (DTC) and Shear sonic log (DTS) are inversely proportional to BI with negative values (−0.89) and (−0.95), respectively. High DTC suggests ductile clay/organic matter, and high values of DTS imply elevated clay content, reducing BI. Studies like [32] [60] [61] confirm that Pearson correlation remains a standard tool in reservoir characterization for quantifying relationships between well logs and petrophysical properties.

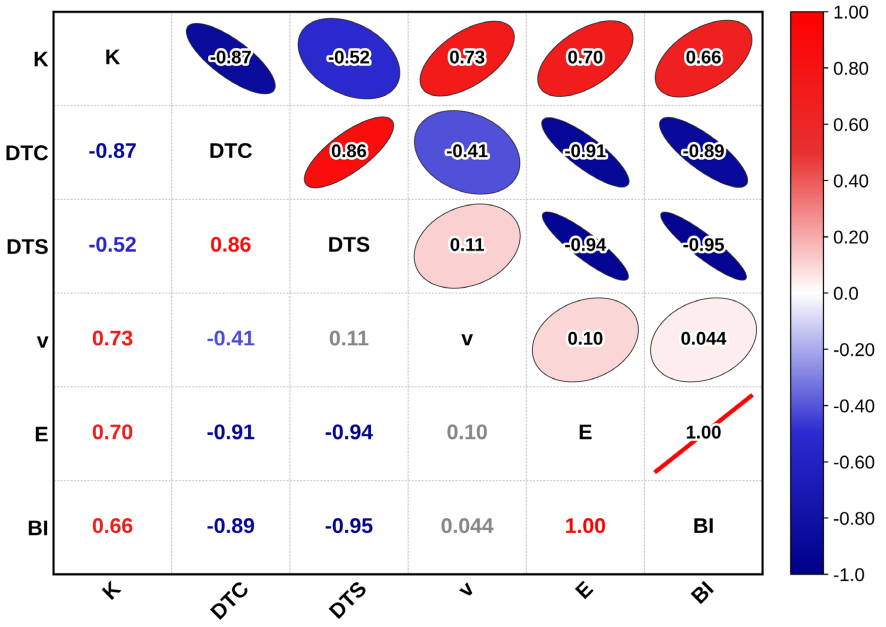

Figure 3. Heatmap of inputs and BI correlation.

3.2.2. Outlier's Detection and Normalization

Data preprocessing was conducted to clean raw logging measurements, mitigate noise and skewness, and enhance model training stability [62]-[64]. Outliers were identified using the Interquartile Range (IQR) method, flagging univariate values beyond the bounds [Q1 - 1.5 × IQR, Q3 + 1.5 × IQR]. Compared to mean-based methods, the IQR approach provides greater resilience to extreme values. To preserve data integrity and maintain the total sample count of 10,852 observations across all wells, flagged outliers were winsorized (capped to the respective upper and lower IQR boundaries) rather than discarded (Figure 4; Figure 5) [65]-[67].

Following outlier treatment, Min-Max normalization was applied to scale all features to a uniform range of [0, 1]. This step eliminates scale discrepancies across heterogeneous wireline logs, ensures balanced feature contributions for distance- and gradient-based algorithms, and improves training convergence [68]-[70].

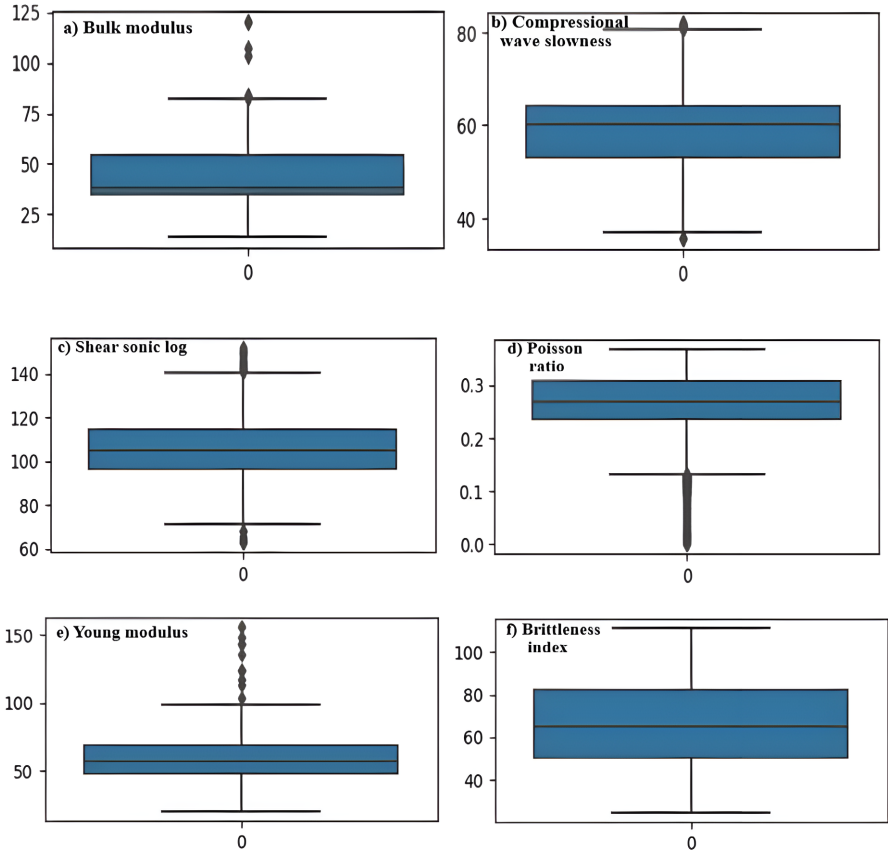

Figure 4. Dataset containing outliers.

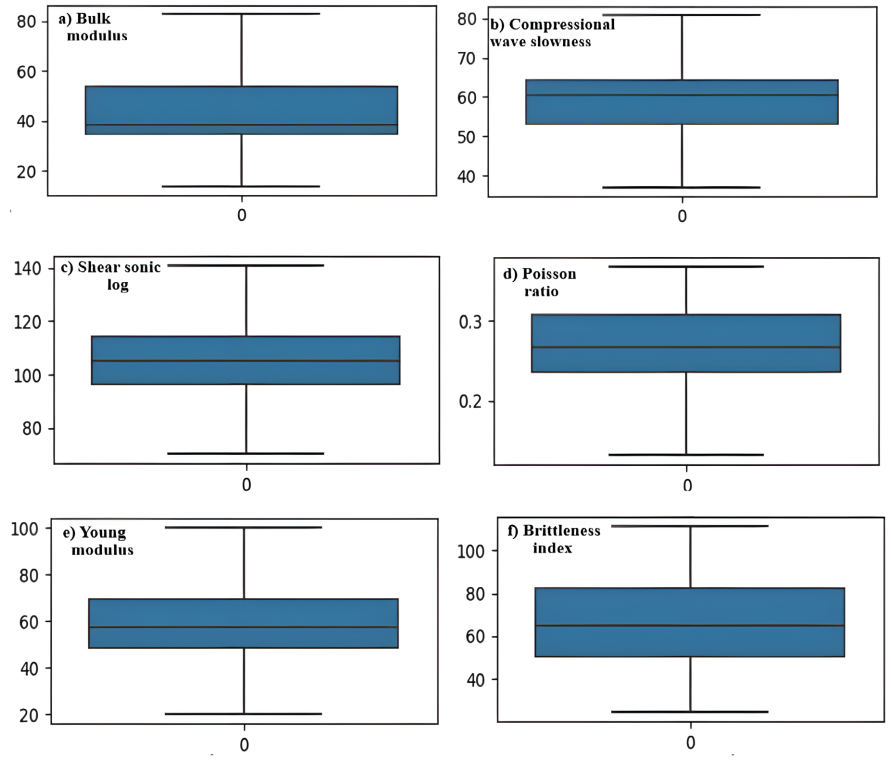

Figure 5. Dataset without outliers.

Critically, to prevent data leakage, both the IQR outlier bounds and the Min-Max scaling parameters were computed strictly from the training dataset (Wells WY1 and WY2). These exact transformation parameters were then applied without modification to the spatial test well (WY3) and the external validation wells (F1 and Y).

All the data were normalized between 0 and 1 using Equation (1).

$$h_{norm} = \frac{h - h_{\min}}{h_{\max} - h_{\min}} \tag{1}$$

where h, $h_{\min}$, $h_{\max}$, h_{norm} are the values to be normalized, the minimum value of h, the maximum value of h and the normalized value of h respectively.

4. Methodology

Three ML algorithms BO-MERF, BPNN, and SVR were utilized to evaluate their robustness and convergence in predicting BI. The models were subsequently applied, with the optimal model selected to estimate BI on unseen data, to ensure the model's validation and generalizability.

4.1. Support Vector Regression (SVR)

The SVR model employs a sophisticated approach involving data mapping and relaxation parameters to perform regression fitting, ensuring that predicted values closely align with actual values for precise forecasting [71] [72]. In support vector regression (SVR), predictions within a designated tolerance range from the actual values are not penalized, allowing for flexibility without compromising accuracy [73]. This approach works by first achieving linear separability in a transformed space, where data previously mapped non-linearly into a high-dimensional space via a kernel function is re-mapped to a more manageable low-dimensional space. This process effectively reduces the discrepancy between expected and actual values, such as in the case of brittleness index predictions in core samples, where this fine-tuned accuracy provides valuable insights into reservoir properties and gas estimates.

4.2. Backpropagation Neural Network (BPNN)

BPNN is widely used in geomechanical modelling for its capability to capture non-linear relationships among input variables [74] [75]. BPNN is a type of multi-layer perceptron (MLP) that uses a gradient descent algorithm to minimize the error between predicted and actual output values by adjusting the weights within the network iteratively. For this study, BPNN was configured with three layers: an input layer, one hidden layer, and an output layer, each containing neurons that activated through a sigmoid function. BPNN used well log and core data as input features during the training process. The weights in each neuron were optimized using back-propagation, where the error from each forward pass was propagated back through the network to adjust the weights, thereby improving prediction accuracy [76] [77]. A mean squared error (MSE) function was employed as the loss function to measure

the discrepancy between predicted and actual stress values, and the model's parameters were tuned using a learning rate and epoch count optimized for convergence. BPNN's capacity to learn complex, nonlinear relationships within data makes it a valuable tool for predicting in-situ stresses in heterogeneous geological formations, although it is computationally demanding and sensitive to overfitting when the dataset is large and high-dimensional [78]. To mitigate overfitting, regularization techniques such as dropout were incorporated, enabling the model to maintain accuracy without excessive reliance on specific data patterns [79].

4.3. Bayesian-Optimized Mixed Effect Random Forest (BO-MERF)

The BO-MERF model offers significant advantages in geomechanical applications, particularly for hydraulic fracturing. BO ensures improved prediction accuracy by identifying optimal hyperparameters, enhancing the model's performance [23]. Additionally, BO reduces computational costs by efficiently exploring the hyperparameter space, eliminating the need for manual tuning [80]. In practical applications, the hybrid model accurately identifies sweet spots with high BI values, optimizing hydraulic fracturing designs for enhanced hydrocarbon recovery [4] [6]. By reducing uncertainty in brittleness predictions, the model minimizes risks associated with targeting unsuitable intervals, ensuring more reliable and efficient resource extraction [81].

4.3.1. Mixed Effect Random Forest (MERF)

Mixed-Effects Random Forest (MERF) extends standard Random Forest (RF) by integrating non-parametric machine learning flexibility with parametric mixed-effects statistical modeling [82] [83]. This hybrid formulation is particularly advantageous for hierarchical, nested datasets—such as wireline logs grouped across multiple wells, geological formations, or geographic basins—where data exhibit both global trends and local group-level variations.

Mathematical Formulation and Algorithmic Workflow

In a mixed-effects framework, response variables are governed by fixed effects (constant parameters across the population) and random effects (group-specific random variables with mean zero) [84]-[86]. Mathematically, the target Brittleness Index (y_{ij}) for the *j*-th observation in the *i*-th group is defined as:

$$y_{ij} = f\left(\boldsymbol{X}_{ij}\right) + \boldsymbol{Z}_{ij}\boldsymbol{b}_{i} + \varepsilon_{ij} \tag{2}$$

where:

$f\left(\boldsymbol{X}_{ij}\right)$ is an ensemble of decision trees (the random forest) representing nonlinear fixed-effects predictions from input covariates $\boldsymbol{X}_{ij}$;

$\boldsymbol{Z}_{ij}$ is the design matrix for group-specific random-effect covariates;

$b_i \sim \mathcal{N}\left(0,\Sigma\right)$ represents group-level random effects assumed to follow a normal distribution with covariance matrix Σ;

$\varepsilon_{ij} \sim \mathcal{N}\left(0,\sigma^2\right)\ \mathcal{N}\left(0,\sigma^2\right)$ represents normally distributed residual errors with variance σ^2.

The MERF algorithm employs an iterative expectation-maximization (EM)-like

optimization scheme [87]:

$$\hat{y}_{ij} = \hat{f}\left(\boldsymbol{X}_{ij}\right) + \boldsymbol{Z}_{ij}\widehat{\boldsymbol{b}_i} \tag{3}$$

During optimization, the forest function $\hat{f}(\cdot)$ is updated by treating the current estimates of random effects as fixed, while the random effects components $\widehat{\boldsymbol{b}_i}$ are updated assuming out-of-bag (OOB) ensemble predictions are valid [33] [88].

Grouping Structure and Generalization to Unseen Wells

To capture geological heterogeneity across the study domain, the model structure was configured as follows:

i) Grouping Variable: Well identifier (Wells WY1, WY2, WY3, F1, and Y).

ii) Random-Effect Covariates ($\boldsymbol{b}_i$): Well-specific intercept adjustments accounting for micro-facies variations, local diagenetic overprints, and regional stress regimes.

iii) Prediction Strategy for Out-of-Sample Wells: For completely unseen wells lacking prior group historical data or core calibration (e.g., Well Y), the unobserved group random effect defaults to zero ($\boldsymbol{b}_i = \boldsymbol{0}$). Under this condition, predictions revert entirely to the robust population-level fixed-effects forest model:

$$y_{ij} = f\left(\boldsymbol{X}_{ij}\right) + \boldsymbol{Z}_{ij}\boldsymbol{b}_i + \varepsilon_{ij} \rightarrow \hat{y}_{ij} = \hat{f}\left(\boldsymbol{X}_{ij}\right) \tag{4}$$

4.3.2. Bayesian Optimization (BO)

BO is a sequential, model-based optimization technique designed to efficiently identify the global optimum of expensive-to-evaluate black-box functions, making it particularly effective for hyperparameter tuning in ML models. BO leverages a probabilistic surrogate model, typically a Gaussian Process ($\mathcal{GP}$), to approximate the objective function $f(x)$ and an acquisition function, such as Expected Improvement (*EI*), to guide the search process. The $\mathcal{GP}$ models the objective function based on observed data points, while the acquisition function balances exploration (sampling uncertain regions) and exploitation (focusing on promising areas) to select the next evaluation point. By minimizing the number of function evaluations required to find the optimal solution, BO significantly reduces computational costs, especially in scenarios where evaluating each configuration is resource-intensive. The surrogate model approximates the objective function $f(x)$ Gaussian Processes (GPs) are commonly used for this purpose due to their flexibility and ability to provide uncertainty estimates [23]. A $\mathcal{GP}$ defines a distribution over functions and is characterized by a mean function $\mu(x)$ and a covariance (kernel) function $k(x, x')$:

$$f(x) \sim \mathcal{GP}\left(\mu(x), k(x, x')\right) \tag{5}$$

where $\mu(x)$ is the mean function, often assumed to be zero or a constant. $k(x, x')$ is the kernel function, which measures the similarity between two inputs x and x'. A common choice is the squared exponential kernel:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|}{2l2}\right) \tag{6}$$

where σ_f^2 is the signal variance and l is the length scale.

Given observed data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ the posterior distribution of $f(x)$ at a new point x_* is Gaussian:

$$f(x_*) \mid \mathcal{D} \mathcal{N}(\mu_*(x_*), \sigma_*^2(x_*)) \tag{7}$$

where $\mu_*(x_*)$ is the posterior mean, $\sigma_*^2(x_*)$ is the posterior variance.

The acquisition function guides the search for the next evaluation point by balancing exploration (sampling uncertain regions) and exploitation (sampling regions likely to improve the objective). A common choice is the Expected Improvement (*EI*):

$$EI(x) = \mathbb{E}\left[\max\left(0, f(x) - f(x^+)\right)\right] \tag{8}$$

where $f(x^+)$ is the best observed value so far. The expectation is taken over the posterior distribution of $f(x)$. For a Gaussian posterior, *EI* can be computed analytically:

$$EI(x) = \left(\mu_*(x) - f(x^+) - \xi\right)\Phi(Z) + \sigma_*(x)\phi(Z) \tag{9}$$

where $Z = \dfrac{\mu_*(x) - f(x^+) - \xi}{\sigma_*(x)}$, Φ and ϕ are the cumulative distribution function (CDF) and probability density function (PDF) of the standard normal distribution, respectively. ξ is a parameter controlling the exploration-exploitation trade-off. The different steps of the algorithm implementation are: (1) Initialize with a few random evaluations of $f(x)$ (2) Fit the $\mathcal{GP}$ surrogate model to the observed data, (3) Use the acquisition function to select the next point x_{n+1}, (4) Evaluate $f(x_{n+1})$ and update the dataset $\mathcal{D}$, (5) Repeat steps 2 - 4 until convergence or a predefined budget is exhausted.

4.3.3. The Hybrid BO-MERF Model Implementation

To predict brittleness in challenging shale reservoirs (like overmatured, overpressured formations), we developed BO-MERF, a hybrid model that strategically combines BO with MERF. First, MERF improves on standard random forests by modeling both systematic trends ("fixed effects") and localized geological variations ("random effects"). This dual approach is crucial for capturing the complex reality of reservoirs. Then, Bayesian Optimization automatically tunes MERF's hyperparameters, squeezing out maximum predictive accuracy. Together, this framework delivers reliable brittleness estimates critical for optimizing hydraulic fracturing. BO has been widely adopted in ML and engineering due to its efficiency in high-dimensional spaces and its ability to handle noisy evaluations. Recent applications in geoscience and reservoir engineering highlight its effectiveness in optimizing complex models, such as predicting brittleness index and CO_2 solubility, with minimal computational overhead [29]. Here are the steps involved in the BO-MERF process:

Step 1: ***Data Preparation:*** Collect and preprocess data (*i.e.*, well logs, seismic

data, geomechanical data) for predicting BI.

Step 2: ***Define the MERF Model***: Specify the fixed effects (input features) and random effects (grouping structure) for the MERF model.

Step 3: ***Set Up Bayesian Optimization***: Define the hyperparameter search space for MERF (*i.e.*, number of trees, maximum depth, random effects parameters).

Step 4: ***Initialize the Surrogate Model***: Use a Gaussian Process or other surrogate model to approximate the objective function (*i.e.*, prediction accuracy).

Step 5: ***Evaluate the Objective Function***: Train the MERF model with the current hyperparameters and evaluate its performance (*i.e.*, using cross-validation).

Step 6: ***Update the Surrogate Model***: Update the surrogate model with the new evaluation results.

Step 7: ***Optimize the Acquisition Function***: Use the acquisition function (*i.e.*, Expected Improvement) to select the next set of hyperparameters to evaluate.

Step 8: ***Iterate Until Convergence***: Repeat steps 5 - 7 until the hyperparameters converge to an optimal solution.

Step 9: ***Train the Final MERF Model***: Train the MERF model with the optimized hyperparameters.

Step 10: ***Predict the Brittleness Index***: Use the trained MERF model to predict the BI for new data. The flowchart for BO-MERF is shown in Figure 6.

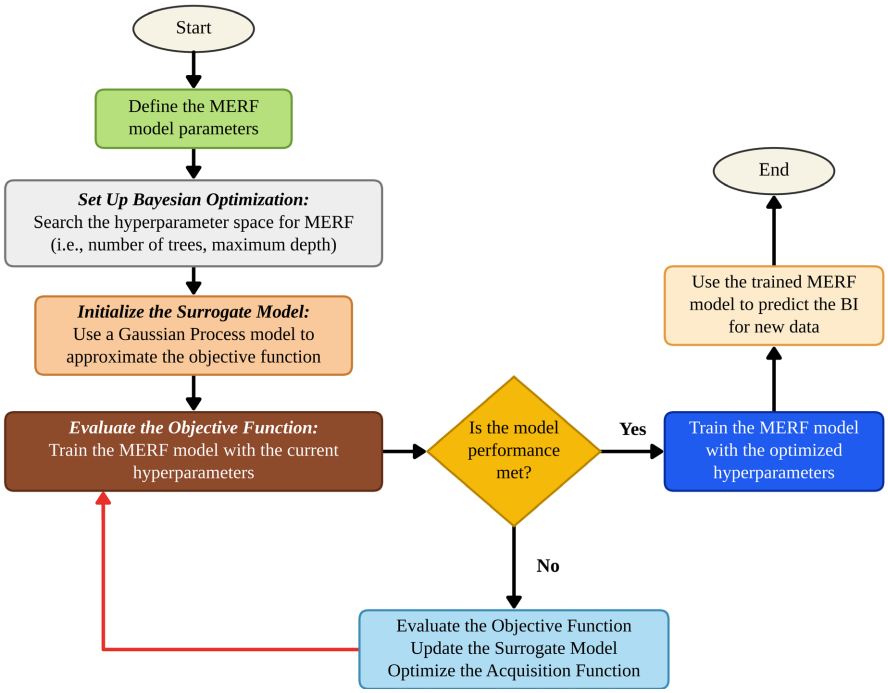

Figure 6. Flow chart of BO-MERF model.

4.4. Model Evaluation Metrics

All the developed models BO-MERF, BPNN, and SVR, were run with Python software version 3.12 in computer running the Windows 10 Professional, operating system 64 bits, equipped with Intel (R) Core (TM) i5-10300H CPU 2.50 GHz. The per-

formance accuracy of the models was evaluated using three key metrics: the mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of determination (R^2), as defined in Equations (10)-(12) [19] [89]. The R^2 metric, which ranges from 0 to 1, quantifies the proportion of variance in the dependent variable explained by the model, with values closer to 1 indicating superior predictive performance. RMSE measures the model's stability and precision by aggregating the magnitude of prediction errors, while MAE provides the average absolute deviation between observed and predicted values, disregarding error direction. A robust regression model is typically characterized by high R^2 values (approaching 1) and low MAE and RMSE values (approaching zero), reflecting accuracy, consistency, and minimal bias [86]. These metrics collectively ensure that the developed models are both reliable and effective for practical applications.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\sigma_i - \sigma'_i\right| \tag{10}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\sigma_i - \sigma'_i\right)^2}{N}} \tag{11}$$

$$R^2 = \left[\frac{\sum_{i=1}^{N}\left(\sigma_i - \bar{\sigma}\right)\left(\sigma'_i - \overline{\sigma'}\right)}{\sqrt{\sum_{i=1}^{N}\left(\sigma_i - \bar{\sigma}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\sigma'_i - \overline{\sigma'}\right)^2}}\right]^2 \tag{12}$$

where σ_i is the measured BI value, $\bar{\sigma}$ is the average value of measured BI, σ'_i is the value of predicted BI, $\overline{\sigma'}$ is the average value of the predicted BI, and N is the total number of data points.

4.5. Hyperparameter Settings

Hyperparameters govern the architecture and training behavior of machine learning algorithms and must be optimized to prevent overfitting or underfitting [28]. To tune the baseline models—Backpropagation Neural Network (BPNN) and Support Vector Regression (SVR)—randomized search was employed to efficiently sample configurations from predefined ranges.

The hyperparameters of the Mixed-Effects Random Forest (MERF) model were exclusively optimized using Bayesian Optimization (BO). Powered by a Gaussian Process surrogate model and an Expected Improvement acquisition function, BO iteratively searches the hyperparameter space to identify high-performing configurations with reduced computational overhead [90]-[95]. To ensure rigorous evaluation without data leakage, the optimization process was guided by Root Mean Square Error (RMSE) as the objective metric, evaluated via a 5-fold cross-validation loop nested strictly within the training dataset (Wells WY1 and WY2). The implementation was executed in Python 3.12 utilizing Scikit-learn and Scikit-optimize, with a fixed random seed (random_state = 42) enforced to guarantee complete reproducibility across runs. Governed by an iteration budget of 100 se-

quential evaluations, the full Bayesian optimization loop completed in 2.08 seconds (representing the total duration of the optimization loop rather than pure single-model training time). The final optimal hyperparameters selected for each model are summarized in Table 2.

Table 2. Hyperparameters tuning of developed models.

Model	Hyperparameters	Range	Optimum value selected
MERF	Number of trees	50 - 500	200
	Maximum depth	5 - 20	10
	Learning rate	0.001 - 0.1	0.01
	Minimum samples split	2 - 10	3
	Minimum samples leaf	1 - 10	2
	Number of iterations	10 - 200	100
	Exploration-exploitation tradeoff	0.5 - 1.5	1.0
	Acquisition function	Expected Improvement (EI) - Probability of Improvement (PI)	Expected Improvement
BPNN	Hidden layers	1 - 3 layers	2 layers
	Neurons per layer	10 - 100 neurons	50 neurons
	Optimizer	Adam, SGD, RMSprop	Adam
	Learning rate	0.0001 - 0.1	0.001
	Batch Size	16 - 128	32
	Epochs	50 - 500	200
SVR	Kernel	Linear, Polynomial, RBF, Sigmoid	RBF
	C (Regularization)	0.1 - 1000	100
	Epsilon (ε)	0.01 - 0.2	0.1

4.6. Validation of the Models

We implemented two complementary validation strategies: the train-test split (Holdout method) and k-fold cross validation to rigorously test our machine learning models for BI prediction in the Sichuan Basin's complex overmature and overpressured shales. First, we applied the Holdout method to data from three key wells (WY1, WY2, and WY3). We reserved 80% of the data for training, teaching the models to map relationships between standard well logs (K, DTC, DTS, E, ν) and BI measurements, while holding back 20% exclusively for testing. This gave us an efficient, real-world assessment of how well the models generalized to unseen data, similar to validation approaches successfully used in reservoir modeling by [96] and porosity/permeability prediction in carbonate reservoirs like [32].

To further validate the models, we implemented k-fold cross-validation (k-fold CV). We divided the three primary wells into k distinct groups, training the model on data from four groups and testing it on the remaining group. This rotation

repeated until every group served as the test set once, ensuring all well data contributed equally to training and evaluation. After validation, we deployed the models to two additional wells with one lacking BI measurement. This approach is validated in geostatistical studies like [97]-[100], proved ideal for capturing spatial variability across wells, simulating real geological contrasts, and confirming model robustness for field deployment. Together, these steps guaranteed thorough evaluation and reliable predictions across diverse well conditions.

5. Results and Discussions

5.1. Model Statistical Analysis

5.1.1. Train-Test Split Analysis

The train-test split (also called the Holdout method) was used to show how our ML models predicted BI. The dataset combined well logs and core measurements from three wells, WY1, WY2, and WY3. We split this data into two parts: 80% for training the models and 20% held back exclusively for testing. During training, the models learned to map relationships between standard well-log curves like K, DTS, DTC, E, and *v*, and the actual BI values. The test set, however, acted as "unseen" data, simulating future wells the models hadn't encountered during training. This let us rigorously evaluate real-world generalization ability, essentially, testing whether the models could reliably predict new data or had just memorized the training patterns. This straightforward method offers a practical balance: it efficiently highlights overfitting while providing a realistic performance estimate. For reservoir modeling applications like ours, where models must generalize to new well locations, this validation approach is both computationally practical and scientifically sound. Recent studies by Andrew, Xu [101] and Mabiala, Cai [32] successfully applied similar methodologies when validating ML predictions for reservoir properties and geomechanical parameters. Our model's performance metrics using this split are detailed in **Table 3**.

Table 3. Performance metrics for BO-MERF with Holdout method.

	R^2	MAE	RMSE
Train	0.9947	0.0162	0.0183
Test	0.9852	0.213	0.459

5.1.2. K-Fold Cross-Validation

K-fold cross-validation (k-fold CV) is a widely adopted best-practice method for evaluating the generalizability, stability, and reliability of ML models, particularly in geosciences and reservoir characterization where data heterogeneity, spatial dependencies, and limited samples are common. This study leverages the same dataset of well-log data and BI measurements from three wells (WY1, WY2, and WY3), the k-fold CV approach partitioned the combined data from all wells into k = 5 mutually exclusive folds of equal size. For each iteration, the model was trained on four folds (80% of the total data) and validated on the remaining fold (20% of the data). This cycle repeated five times, ensuring every fold served as the validation set exactly

once, while the other four folds were used for training (Figure 7).

This k-fold cross-validation method exposes models to diverse well data, crucial for robust geological characterization (*i.e.*, lithology prediction). It effectively assesses spatial heterogeneity and provides stable performance metrics by aggregating across folds, simulating real-world blended data scenarios. While standard k-fold can be optimistic due to spatial autocorrelation, modified spatial k-fold addresses this. It's vital for hyperparameter tuning and model selection. Although it may slightly overestimate error, this is outweighed by its benefits in handling spatial heterogeneity and ensuring reliable evaluation for geoscience [100] [102]-[107]. The consolidated results of this evaluation are documented in Table 4.

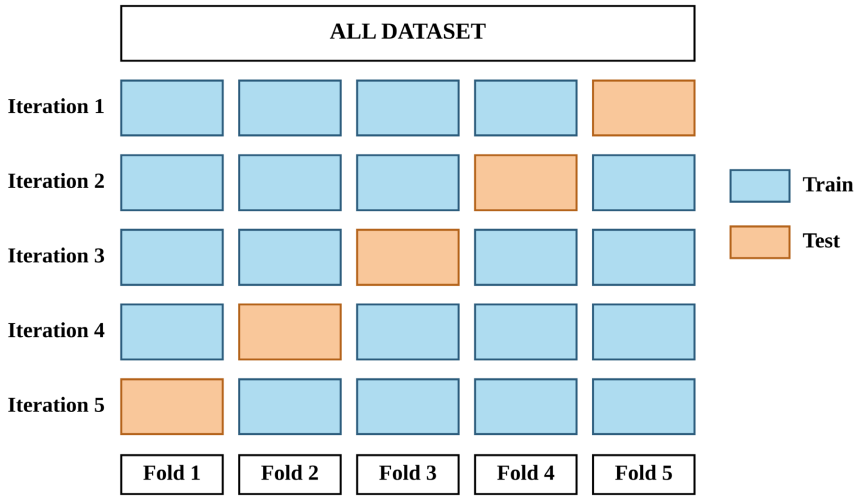

Figure 7. Schematic process of the k-fold CV method for BI prediction.

Table 4. Performance Metrics for BO-MERF with 5-fold CV method.

K-folds	R^2	MAE	RMSE
k = 1	0.9820	0.24	0.555
k = 2	0.9760	0.18	0.523
k = 3	0.9750	0.22	0.595
k = 4	0.9680	0.255	0.580
k = 5	0.979	0.23	0.594
Average	0.9756	0.225	0.569

5.2. Statistical Error Analysis

Figure 8 and Figure 9 show predicted vs. actual BI scatter plots for training and testing, respectively, with the diagonal indicating perfect fit. BO-MERF achieved near-perfect training fit with $R^2 = 0.9947$ and maintained excellent testing performance with $R^2 = 0.9756$, demonstrating high accuracy and generalizability. BPNN showed good training fit but significant testing deviation, indicating overfitting. SVR had the weakest training fit and poorest testing performance, with severe overfitting. BO-MERF's superior results, closest to the diagonal in

both phases, confirm its robustness and best-in-class performance, attributable to efficient bayesian hyperparameter tuning and MERF's ability to capture data relationships.

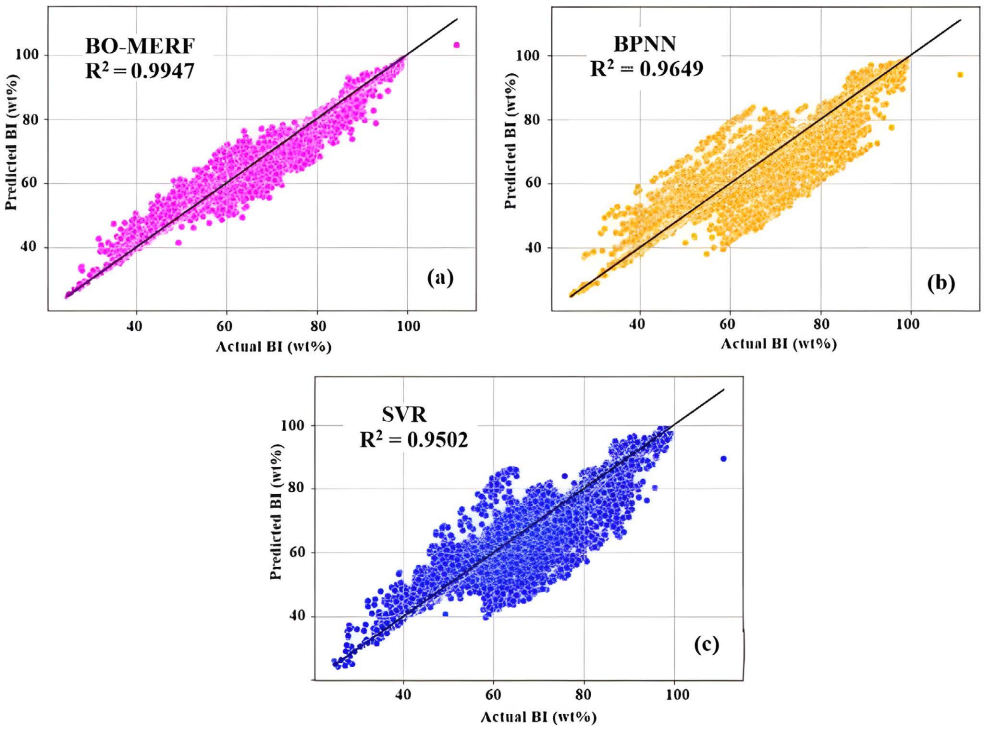

Figure 8. Scatter plot for training data.

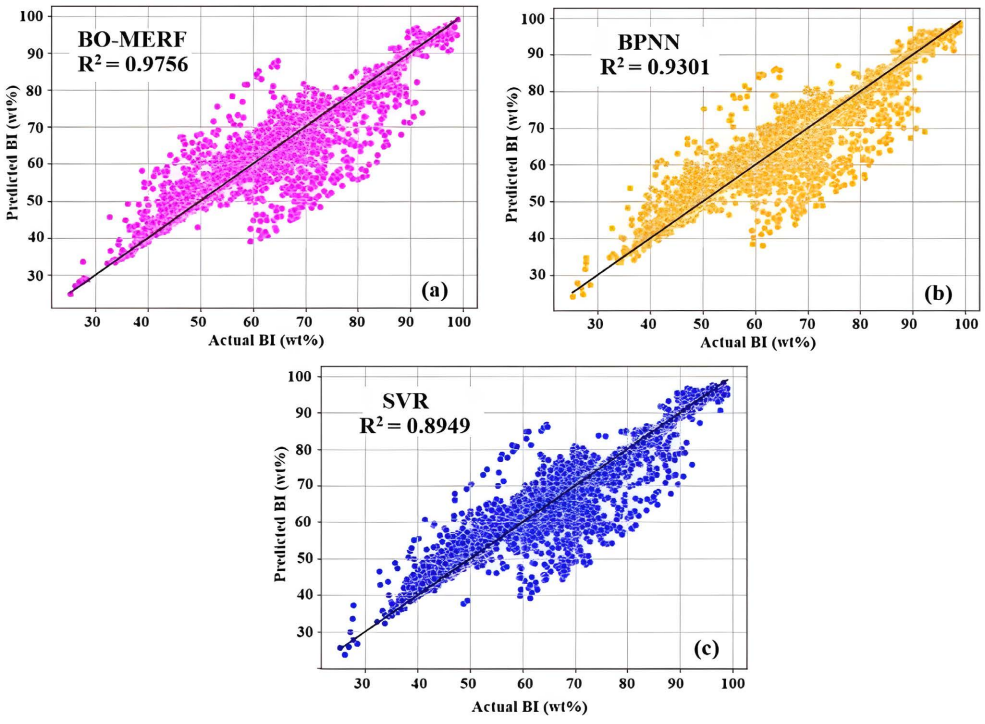

Figure 9. Scatter plot for testing data.

BO-MERF outperformed BPNN and SVR in training and testing, achieving the lowest RMSE and MAE (Figure 10 and Figure 11), indicating robustness against noise and outliers. While BPNN and SVR showed significant test performance degradation due to overfitting, BO-MERF maintained high accuracy and stability, demonstrating superior generalization. As shown in Table 5, BO-MERF runs faster (2.08 s) than BPNN (4.32 s) and SVR (6.02 s) due to smarter tuning via bayesian optimization and its MERF component handling spatial dependencies. This combination, lacking in BPNN (tuning/interpretability issues) and SVR (scalability/ignoring spatial data), makes BO-MERF a superior, reliable tool for shale gas reservoir characterization.

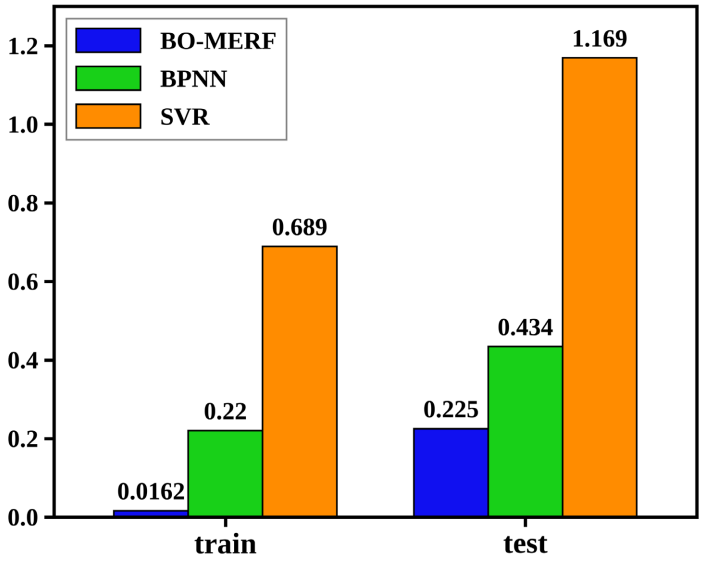

Figure 10. Comparisons of MAE for developed models.

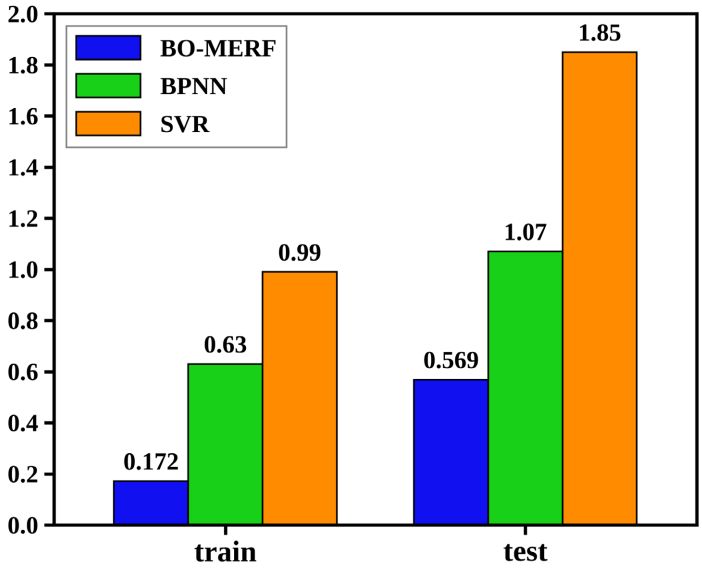

Figure 11. Comparisons of RMSE for developed models.

Table 5. Statistical performance indicators for developed models.

Model	R^2		MAE		RMSE		Computational
	Train	Test	Train	Test	Train	Test	time (sec)
BO-MERF	0.9947	0.9756	0.0162	0.225	0.0172	0.569	2.08
BPNN	0.9649	0.9301	0.22	0.434	0.63	1.07	4.32
SVR	0.9502	0.8949	0.689	1.169	0.99	1.85	6.02

5.3. Model Comparison

The comparison between the developed models to estimate the BI was performed through the Taylor diagram and violin plots. The Taylor diagram is a powerful graphical tool used to evaluate the performance of models by comparing their statistical agreement with observed data. It provides a concise summary of three key metrics: the correlation coefficient (R), the root-mean-square error (RMSE), and the standard deviation (σ). The diagram plots these metrics on a polar coordinate system, where the radial distance from the origin represents the standard deviation, the azimuthal angle represents the correlation coefficient, and the distance from the reference point (observed data) represents the centered RMSE [108]. It has been widely applied in climate modeling [109], hydrology [110], and geosciences [72] [86], providing a clear and intuitive way to assess model skill. **Figure 12** revealed that the BO-MERF model outperformed others, positioning closest to the reference point, indicating superior performance with an optimal balance between R, RMSE, and σ. It achieved the highest R^2 reflecting a strong correlation between predicted and observed BI values, along with low RMSE and a standard deviation close to the reference, confirming accurate and well-distributed predictions. The BPNN model ranked second, with slightly lower R and higher RMSE but maintained a good balance. The SVR model showed comparable RMSE to BPNN but had a lower R-value and higher standard deviation, indicating less reliability with the lowest R, highest RMSE, and a significantly higher standard deviation, reflecting larger errors and wider prediction dispersion. The analysis validated BO-MERF as the most precise and accurate for BI prediction, consistently aligning with actual data. According to R^2, the performance order is BO-MERF > BPNN > SVR.

On the other hand, violin plots were employed to evaluate model performance in predicting BI. The BO-MERF model exhibited superior predictive accuracy, as indicated by the precise alignment of its median (denoted by white dots within the violin plot) with the observed experimental data. As illustrated in **Figure 13**, this model also demonstrated the closest agreement in predicting both lower and upper percentiles (5th and 95th, represented by thin black lines), as well as the interquartile range (25th and 75th percentiles, marked by thicker lines), outperforming the alternative models. The BPNN model exhibited a comparable level of performance, closely mirroring the accuracy of the BO-MERF model across all statistical indicators. Conversely, the SVR model tended to overestimate the lower

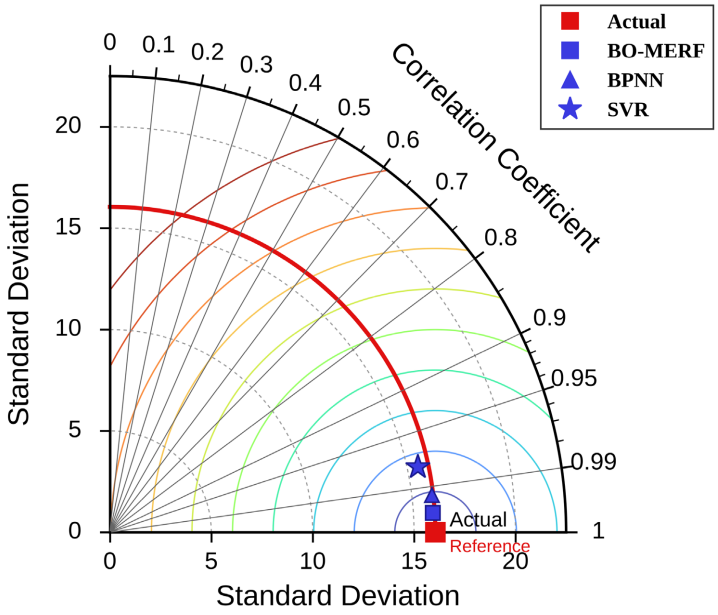

Figure 12. Taylor diagram for BI estimation.

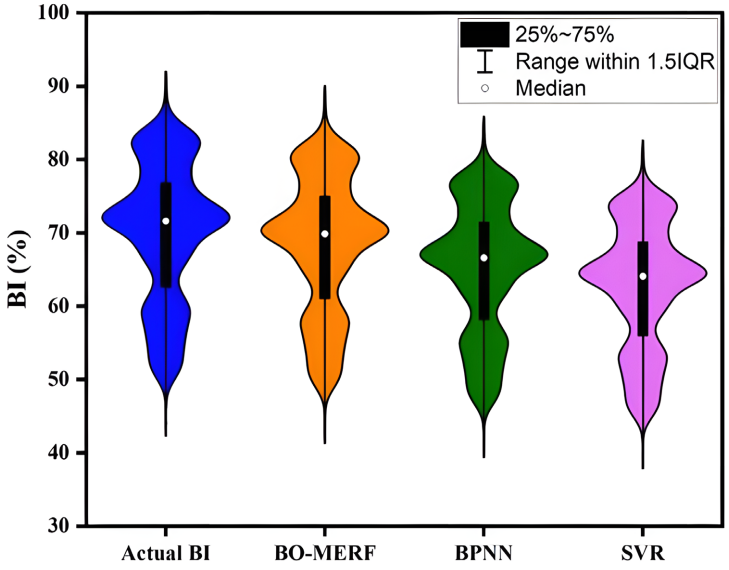

Figure 13. Violin plots for BI estimation.

tail while underestimating the upper tail of the BI range. Regarding the probability density function (PDF), as represented in the violin plot, the BO-MERF model's PDF demonstrated the highest alignment with the actual experimental BI data distribution, followed by BPNN, whereas the SVR model displayed the least agreement. The model performance ranking is BO-MERF > BPNN > SVR, aligning with the previous ones discussed above. The robust performance of the BO-MERF model across diverse formations, including challenging environments characterized by complex lithological variations, further validates its capability to handle intricate scenarios. BI estimation in such settings is particularly difficult due to the heterogeneity of rock properties and the influence of mineralogical composi-

tion. The BO-MERF model's ability to generalize well and maintain high accuracy under these conditions highlights its superiority over traditional and other ML approaches. This makes it a reliable tool for predicting BI in shale gas reservoirs, where accurate estimation is critical for hydraulic fracturing design and sweet spot identification.

5.4. SHAP Analysis

SHAP (Shapley Additive Explanations) is an interpretability framework based on cooperative game theory, designed to quantify the contribution of individual features in ML predictions through Shapley values [111]. This method enhances model transparency, enabling researchers to identify influential features, detect biases, and validate model reliability [112] [113]. Unlike other interpretability techniques, SHAP captures both individual feature effects and interactions, making it particularly useful for complex, nonlinear models [114]. SHAP values can be visualized through several methods (*i.e.*, summary plots, bar plots, dependence plots), each offering unique insights into model behavior. The summary plot aggregates SHAP values across all instances, where each dot represents a data point. Features are ordered vertically by global importance, and the horizontal position reflects the SHAP value magnitude (positive/negative impact). Color intensity indicates the feature value (red: high, blue: low). Additionally, SHAP bar plots rank features based on their mean absolute SHAP values, offering a straightforward measure of overall feature importance. Unlike beeswarm or dependence plots, bar plots focus solely on contribution magnitude rather than directionality, making them useful for summarizing key predictors but less effective in revealing feature interactions.

Figure 14 and **Figure 15** demonstrate Young's Modulus (E) as the dominant predictor of BI in the BO-MERF model. The SHAP summary plot (**Figure 14**) reveals a strong positive correlation: high E values (red dots) increase BI, while low E values decrease it. This dominance is quantitatively confirmed by the SHAP bar plot (**Figure 15**), where E exhibits the highest mean impact (9.5), significantly exceeding other features (mean SHAP < 3). While parameters like bulk modulus (K), compressional slowness (DTC), and shear slowness (DTS) show weaker positive influences on BI, Poisson's ratio (ν) exhibits an inverse relationship—lower ν values increase BI.

Young's modulus preeminence reflects fundamental rock mechanics: High E signifies a stiff rock capable of storing elastic strain energy, which is catastrophically released as brittle fractures during hydraulic fracturing. This role is amplified in the Longmaxi shale due to its unique geology—exceptionally high biogenic quartz content (>40% - 50%) and advanced thermal maturity transform organic matter into rigid graphitic carbon, creating a brittle fabric enhanced by natural fractures [115]-[122]. Consequently, quartz-rich intervals maximize fracture network complexity during stimulation.

The BO-MERF model's accuracy validates the dual control of E (dominant in brittle silica layers) and ν (more influential in ductile clay zones) on BI, reflecting

core studies of shale heterogeneity. Its successful use of random effects further confirms the impact of vertical heterogeneity (interbedded lithofacies) on fracturability, aligning E's SHAP-derived dominance with the Longmaxi Formation's intrinsic rock properties.

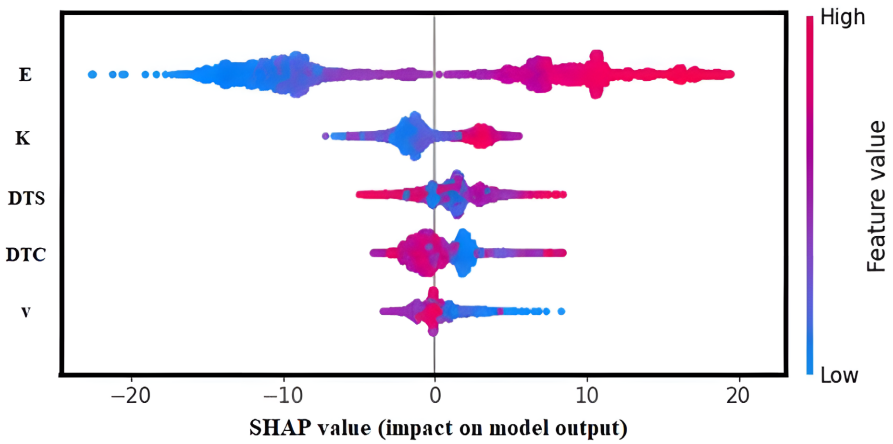

Figure 14. Summary plot showing the inputs influence on BI estimation.

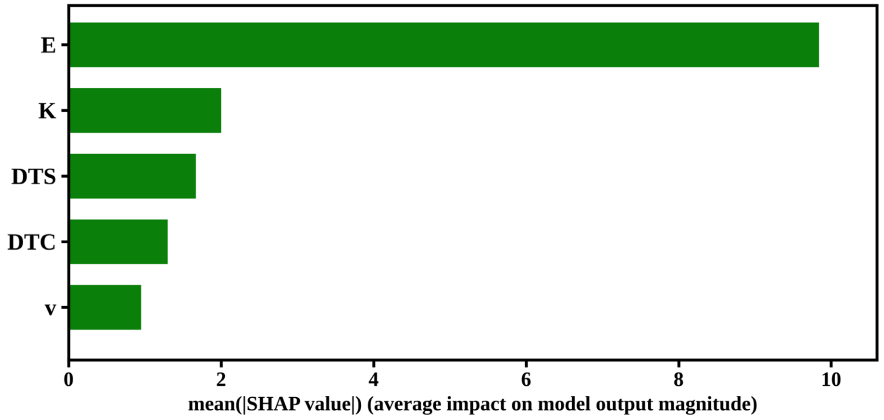

Figure 15. Bar plots showing the inputs importance for BI estimation.

5.5. Model Validation

Following successful training and validation on Weiyuan field wells (WY1 - WY3), the BO-MERF model—utilizing input parameters including K, DTC, DTS, E, and ν—was extended to two unseen external wells for rigorous performance evaluation: Well F1 (quantitative blind test) from the Fuling shale gas field and Well Y (qualitative cross-basin generalization) from the Marcellus Shale. To validate model accuracy, measured BI data for well F1 was masked during prediction. The model demonstrated exceptional predictive performance with strong agreement against field data, achieving an R^2 of 0.9825, an RMSE of 0.035, and an MAE of 0.014 (Figure 16). These near-zero error metrics and an R^2 approaching unity confirm the model's robustness and local transferability. Given the complete absence of core-derived BI ground truth for the Marcellus Shale, Well Y served as a qualitative generalization test. Rather than relying on quantitative error metrics,

the predicted BI profile was evaluated against established regional geomechanical ranges for the formation (typically 0.3 to 0.85, with values exceeding 0.5 favoring efficient hydraulic fracturing). The predicted BI profile demonstrated stable, geologically consistent values across the entire depth interval (Figure 17), underscoring the model's reliability for cross-basin application [123]-[125].

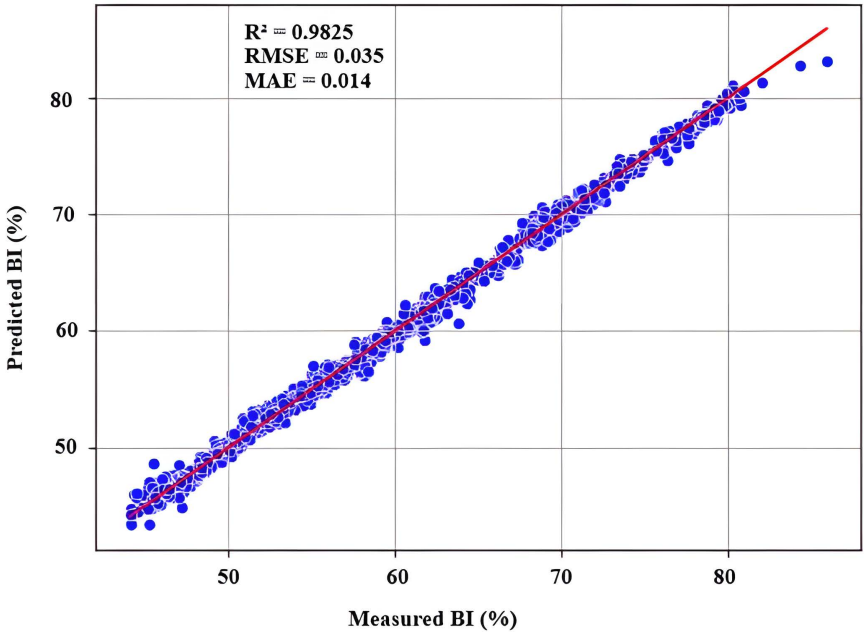

Figure 16. Cross plots between actual and predicted BI for well F1.

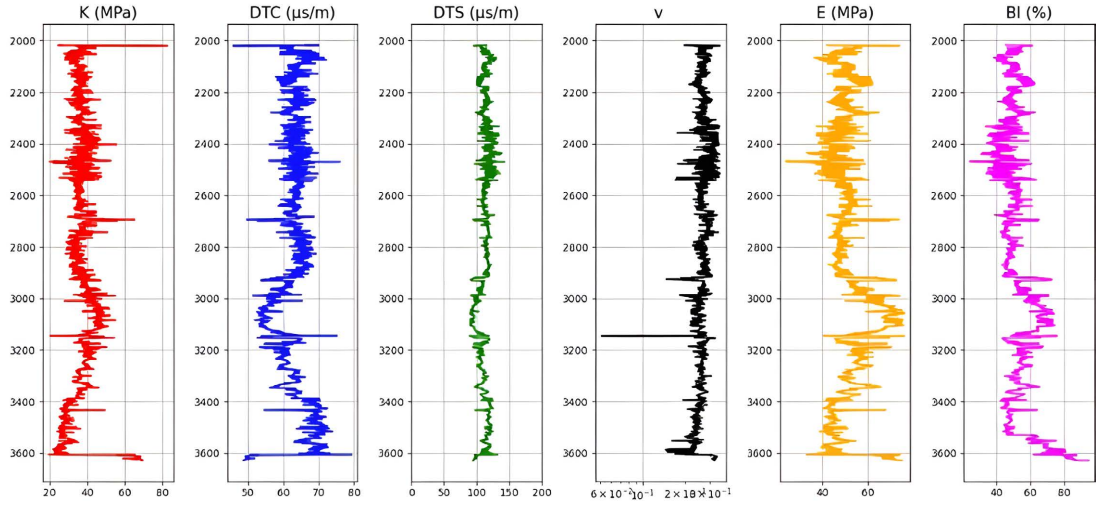

Figure 17. Data inputs with predicted BI values using BO-MERF from well Y.

5.6. Comparison of BO-MERF with Previous Studies

Table 6 compares the performances of the proposed BO-MERF model against two established models, random forest optimized with particle swarm optimization (RF-PSO) and artificial neural network (ANN), for BI estimation. The evaluation metrics include R^2 and RMSE for both training and testing datasets. The

BO-MERF model outperformed the ANN and RF-PSO models developed by Mustafa, Tariq [126] and Nadege, Jiang [19], respectively. During the training phase, the BO-MERF model achieved the highest R^2 value of 0.9947, followed by RF-PSO (R^2 = 0.9934) and ANN (R^2 = 0.96). Similarly, during testing, BO-MERF maintained superior performance with the highest R^2 value of 0.9756, surpassing ANN (R^2 = 0.962) and RF-PSO (R^2 = 0.9533) (**Figure 18**). The BO-MERF model also demonstrated significantly lower error metrics. For training, BO-MERF recorded the lowest RMSE of 0.0172, compared to ANN (RMSE = 2.27) and RF-PSO (RMSE = 4.6327). During testing, BO-MERF's RMSE (0.569) remained substantially lower than RF-PSO (15.53), compared to ANN (2.35); highlighting its robustness in generalizing to unseen data. This superior performance confirms BO-MERF as the most reliable model for BI estimation. The enhanced accuracy of BO-MERF stems from its hybrid framework: BO efficiently tunes hyperparameters and selects optimal features, while MERF captures complex nonlinear relationships and accounts for both fixed and random effects in heterogeneous geomechanical datasets. This synergy minimizes overfitting and ensures consistent

Table 6. Comparison of BO-MERF with previous studies for predicting BI.

Models	R^2		RMSE		Field	Country	References
	Train	Test	Train	Test			
BO-MERF	0.9947	0.9756	0.0172	0.569	Sichuan basin	China	This study
RF-PSO	0.9934	0.9533	4.6327	15.53	Sichuan basin	China	Nadege, Jiang *et al.* [19]
ANN	0.96	0.962	2.27	2.35	Arabian Peninsula	Middle East	[126]

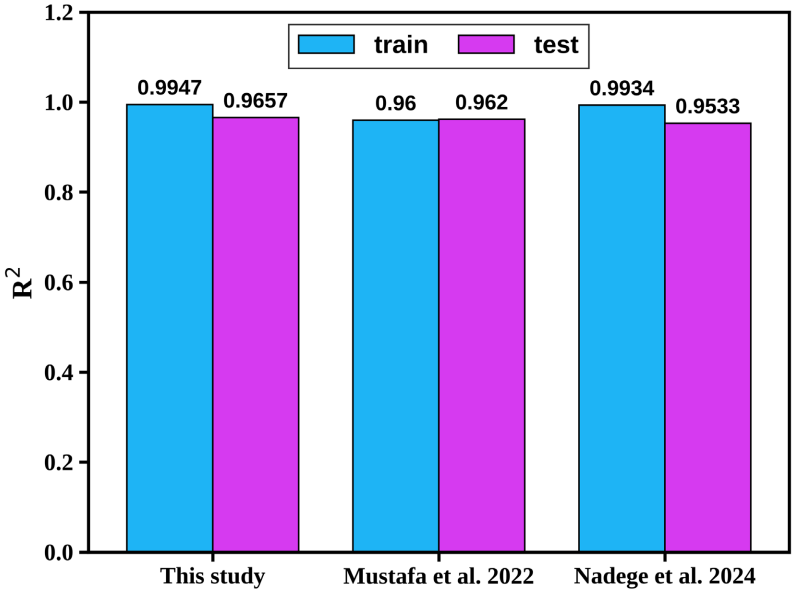

Figure 18. R^2 comparison of BO-MERF model (this study) with ANN [126] and RF-PSO [19] for BI prediction.

error reduction across training and testing phases. In contrast, ANN struggles with high variance in error metrics, and RF-PSO's reliance on particle swarm optimization leads to suboptimal parameter tuning, resulting in poorer generalization. The overall ranking of model performance is BO-MERF > ANN > RF-PSO, with BO-MERF emerging as the most accurate and generalizable choice for BI prediction.

5.7. Limitations and Recommendations of the Work

Key limitations include the model's training solely on vertical well data, where accuracy may drop in inclined wells due to shale anisotropy; its restriction to four wells in Sichuan Basin's Weiyuan/Fuling fields and Longmaxi Formation depths, requiring calibration for broader applicability; confined geomechanical ranges with poor interpretability/transferability from insufficient data; and a narrow BI focus, excluding critical fracturing parameters (*i.e.*, stress anisotropy, fracability), limiting holistic pay-zone identification.

To address these gaps, future work must expand datasets (diverse wells/formations, 3D seismic, stress anisotropy), validate against global benchmarks (*i.e.*, Tarim and Ordos Basins), and integrate cross-disciplinary data (production, lithology), essential multi-disciplinary enhancements for transferable, field-scale fracturing optimization and sweet-spot mapping.

6. Conclusions

This study presents the novel Bayesian Optimization-Mixed Effects Random Forest (BO-MERF) hybrid model to predict the brittleness index (BI), a critical geomechanical parameter indicating rock fracturability, for optimizing hydraulic fracturing in the Weiyuan gas fields of China's Sichuan Basin. The BO-MERF model was applied to predict BI within the Wufeng-Longmaxi Formation using the input parameters E, K, DTC, DTS, and *v*. The main findings are listed below:

1) The BO-MERF model achieved superior performance in BI prediction compared to BPNN and SVR, with R^2 values of 0.9947 for training and 0.9756 for testing, RMSE values of 0.0172 and 0.569, and MAE values of 0.0162 and 0.225, respectively. The models ranked in order of performance as BO-MERF > BPNN > SVR, demonstrating the highest R^2 and the lowest error metrics. This confirms that BO-MERF can predict better the BI of the Sichuan basin.

2) The BO-MERF model was evaluated for performance and generalization using well F1 from the Fuling shale field as a quantitative blind test, and well Y from the Marcellus shale field as a qualitative cross-basin test. The model achieved high predictive accuracy on Well F1 while maintaining reliable performance on Well Y despite the lack of core-derived BI reference data. Subsequent SHAP analysis identified Young's modulus as the most influential feature driving the model's BI predictions.

3) Additionally, the performance of the BO-MERF model was evaluated against previously developed models, including RF-PSO and ANN, with the BO-MERF

model demonstrating superior results compared to these models.

4) The BI alone is insufficient for fracturing decisions; complementary parameters like fracture toughness, tensile strength, anisotropy stress, fracability index, and in-situ stress must be integrated to optimize fracture placement in shale gas reservoirs. Therefore, the BO-MERF model has the potential to be applied to predict additional hydraulic fracturing properties and reservoir parameters, not only in the Sichuan Basin but also in other regions globally.

Data Availability Statement

Data is available upon reasonable request from corresponding author.

Author Contributions

Conceptualization, Yao Hugues Roland Konan; Methodology, Yao Hugues Roland Konan; Software, Yao Hugues Roland Konan; Validation, Kouame Jean-Medard Bouaffou, N'guessan Donald Ahoure; Formal analysis, Yao Hugues Roland Konan, Kouame Jean-Medard Bouaffou, N'guessan Donald Ahoure; Investigation, Yao Hugues Roland Konan; Data curation, Yao Hugues Roland Konan; Writing—original draft preparation, Yao Hugues Roland Konan; Writing—review and editing, Yao Hugues Roland Konan, Kouame Jean-Medard Bouaffou, N'guessan Donald Ahoure, Zéli Bruno Digbehi; Visualization, Kouame Jean-Medard Bouaffou; Supervision, Zéli Bruno Digbehi.

Conflict of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Newell, R., *et al.* (2021) Global Energy Outlook 2021: Pathways from Paris. Resources for the Future, 39 p.

[2] Nalley, S. and LaRose, A. (2022) Annual Energy Outlook 2022 (AEO2022). Energy Information Agency, 23 p.

[3] Cai, F., Ma, Y. and Jin, Z. (2022) Annual Report on China's Petroleum, Gas and New Energy Industry (2021). Springer.

[4] Rickman, R., Mullen, M., Petre, E., Grieser, B. and Kundert, D. (2008) A Practical Use of Shale Petrophysics for Stimulation Design Optimization: All Shale Plays Are Not Clones of the Barnett Shale. *SPE Annual Technical Conference and Exhibition*, Denver, 21-24 September 2008, SPE-115258-MS. https://doi.org/10.2118/115258-ms

[5] Jarvie, D.M., Hill, R.J., Ruble, T.E. and Pollastro, R.M. (2007) Unconventional Shale-Gas Systems: The Mississippian Barnett Shale of North-Central Texas as One Model for Thermogenic Shale-Gas Assessment. *AAPG Bulletin*, **91**, 475-499. https://doi.org/10.1306/12190606068

[6] Wang, F.P. and Gale, J.F. (2009) Screening Criteria for Shale-Gas Systems. *Gulf Coast Association of Geological Societies Transactions*, **59**, 779-793.

[7] Rojas, L.F., Peña, Y.Q. and Carrillo, Z.H. (2016) Brittleness Analysis: A Methodology

to Identify Sweet Spots in Shale Gas Reservoirs. *SPE Argentina Exploration and Production of Unconventional Resources Symposium*, Buenos Aires, 1-3 June 2016, SPE-180955-MS. https://doi.org/10.2118/180955-ms

[8] Zoback, M.D. (2010) Reservoir Geomechanics. Cambridge University Press.

[9] Shi, X., Liu, G., Cheng, Y., Yang, L., Jiang, H., Chen, L., *et al.* (2016) Brittleness Index Prediction in Shale Gas Reservoirs Based on Efficient Network Models. *Journal of Natural Gas Science and Engineering*, **35**, 673-685. https://doi.org/10.1016/j.jngse.2016.09.009

[10] Khanjar, H. (2024) Applications of Machine Learning in Sweet-Spots Identification: A Review. *SPE Journal*, **29**, 6623-6639. https://doi.org/10.2118/223620-pa

[11] Tang, J., Fan, B., Xiao, L., Tian, S., Zhang, F., Zhang, L., *et al.* (2020) A New Ensemble Machine-Learning Framework for Searching Sweet Spots in Shale Reservoirs. *SPE Journal*, **26**, 482-497. https://doi.org/10.2118/204224-pa

[12] Zhong, R., Johnson, R.L. and Chen, Z. (2020) Using Machine Learning Methods to Identify Coal Pay Zones from Drilling and Logging-While-Drilling (LWD) Data. *SPE Journal*, **25**, 1241-1258. https://doi.org/10.2118/198288-pa

[13] Rahimzadeh Kivi, I., Zare-Reisabadi, M., Saemi, M. and Zamani, Z. (2017) An Intelligent Approach to Brittleness Index Estimation in Gas Shale Reservoirs: A Case Study from a Western Iranian Basin. *Journal of Natural Gas Science and Engineering*, **44**, 177-190. https://doi.org/10.1016/j.jngse.2017.04.016

[14] Chan, S., Hassan, A., Humphrey, J.D., *et al.* (2020) Evaluation of Brittleness Index Based on Mineral Compositions Prediction Using Artificial Neural Network. *ARMA/DGS/SEG International Geomechanics Symposium*, Virtual, 3-5 November 2020.

[15] Ahmadov, J. (2021) Utilizing Data-Driven Models to Predict Brittleness in Tuscaloosa Marine Shale: A Machine Learning Approach. *SPE Annual Technical Conference and Exhibition*, Dubai, 21-23 September 2021, SPE-208628-STU. https://doi.org/10.2118/208628-stu

[16] Yuan, Q., Chen, Y., Yao, Y., Fang, Y., Chen, Z. and Su, H. (2024) Machine Learning-Based Sweet Spot Prediction Method for Canada Tight Sandstone Gas Reservoir. *SPE/AAPG/SEG Unconventional Resources Technology Conference*, Houston, 17-19 June 2024, URTEC-4025206-MS. https://doi.org/10.15530/urtec-2024-4025206

[17] Algarhy, A., Adel Gabry, M., Farid Ibrahim, A., Gharieb, A. and Darraj, N. (2024) Optimizing Development Strategies for Unconventional Reservoirs of Abu Roash Formation in the Western Desert of Egypt. *SPE/AAPG/SEG Unconventional Resources Technology Conference*, Houston, 17-19 June 2024, URTEC-4044905-MS. https://doi.org/10.15530/urtec-2024-4044905

[18] Aminaho, E.N. (2024) Caprock Integrity Evaluation for Geosequestration of CO_2 in Depleted Petroleum Reservoirs. Ph.D. Thesis, Robert Gordon University.

[19] Nadege, M.N., Jiang, S., Mwakipunda, G.C., Kouassi, A.K.F., Harold, P.K. and Hugues Roland, K.Y. (2024) Brittleness Index Prediction Using Modified Random Forest Based on Particle Swarm Optimization of Upper Ordovician Wufeng to Lower Silurian Longmaxi Shale Gas Reservoir in the Weiyuan Shale Gas Field, Sichuan Basin, China. *Geoenergy Science and Engineering*, **233**, Article ID: 212518. https://doi.org/10.1016/j.geoen.2023.212518

[20] Mwakipunda, G.C., Komba, N.A., Ayimadu, E.T. and Yu, L. (2025) Model Development for Brittleness Index Estimation and Depth Determination in Hydraulic Fracturing Operations in Shale Gas Reservoirs Using Machine Learning Algorithms. *SPE Journal*, **30**, 4409-4430. https://doi.org/10.2118/228285-pa

[21] Ye, Y., Tang, S., Xi, Z., Jiang, D. and Duan, Y. (2022) A New Method to Predict Brit-

tleness Index for Shale Gas Reservoirs: Insights from Well Logging Data. *Journal of Petroleum Science and Engineering*, **208**, Article ID: 109431. https://doi.org/10.1016/j.petrol.2021.109431

[22] Yasin, Q., Du, Q., Sohail, G.M. and Ismail, A. (2018) Fracturing Index-Based Brittleness Prediction from Geophysical Logging Data: Application to Longmaxi Shale. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, **4**, 301-325. https://doi.org/10.1007/s40948-018-0088-4

[23] Snoek, J., Larochelle, H. and Adams, R.P. (2012) Practical Bayesian Optimization of Machine Learning Algorithms. arXiv: 1206.2944.

[24] Awal, M.A., Masud, M., Hossain, M.S., Bulbul, A.A., Mahmud, S.M.H. and Bairagi, A.K. (2021) A Novel Bayesian Optimization-Based Machine Learning Framework for COVID-19 Detection from Inpatient Facility Data. *IEEE Access*, **9**, 10263-10281. https://doi.org/10.1109/access.2021.3050852

[25] Dewancker, I., McCourt, M. and Clark, S. (2016) Bayesian Optimization for Machine Learning: A Practical Guidebook. arXiv: 1612.04858.

[26] Injadat, M., Salo, F., Nassif, A.B., Essex, A. and Shami, A. (2018) Bayesian Optimization with Machine Learning Algorithms towards Anomaly Detection. 2018 *IEEE Global Communications Conference* (*GLOBECOM*), Abu Dhabi, 9-13 December 2018, 1-6. https://doi.org/10.1109/glocom.2018.8647714

[27] Wu, J., *et al.* (2019) Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, **17**, 26-40.

[28] Bergstra, J. and Bengio, Y. (2012) Random Search for Hyper-Parameter Optimization. *The Journal of Machine Learning Research*, **13**, 281-305.

[29] Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. and de Freitas, N. (2016) Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, **104**, 148-175. https://doi.org/10.1109/jproc.2015.2494218

[30] Zhang, W., Wu, C., Zhong, H., Li, Y. and Wang, L. (2021) Prediction of Undrained Shear Strength Using Extreme Gradient Boosting and Random Forest Based on Bayesian Optimization. *Geoscience Frontiers*, **12**, 469-477. https://doi.org/10.1016/j.gsf.2020.03.007

[31] Mwakipunda, G.C., Nafouanti, M.B., Ibrahim, A. and Yu, L. (2025) Estimating the Hydrogen Adsorption in Depleted Shale Gas Reservoirs for Kerogens in Underground Hydrogen Storage Using Machine Learning Algorithms. *Fuel*, **388**, Article ID: 134534. https://doi.org/10.1016/j.fuel.2025.134534

[32] Mabiala, A.P., Cai, Z., Kouassi, A.K.F., Zhang, H., Mwakipunda, G.C. and Mahamadou, A.S. (2025) Integrating Advanced Machine Learning Models for Accurate Prediction of Porosity and Permeability in Fractured and Vuggy Carbonate Reservoirs: Insights from the Tarim Basin, Northwestern, China. *SPE Journal*, **30**, 3307-3333. https://doi.org/10.2118/226198-pa

[33] Krennmair, P. and Schmid, T. (2022) Flexible Domain Prediction Using Mixed Effects Random Forests. *Journal of the Royal Statistical Society Series C*: *Applied Statistics*, **71**, 1865-1894. https://doi.org/10.1111/rssc.12600

[34] Dai, J., Zou, C., Dong, D., Ni, Y., Wu, W., Gong, D., *et al.* (2016) Geochemical Characteristics of Marine and Terrestrial Shale Gas in China. *Marine and Petroleum Geology*, **76**, 444-463. https://doi.org/10.1016/j.marpetgeo.2016.04.027

[35] Zou, C., Dong, D., Wang, S., Li, J., Li, X., Wang, Y., *et al.* (2010) Geological Characteristics and Resource Potential of Shale Gas in China. *Petroleum Exploration and Development*, **37**, 641-653. https://doi.org/10.1016/s1876-3804(11)60001-3

[36] Liu, S., Yang, Y., Deng, B., Zhong, Y., Wen, L., Sun, W., *et al.* (2021) Tectonic Evolution of the Sichuan Basin, Southwest China. *Earth-Science Reviews*, **213**, Article ID: 103470. https://doi.org/10.1016/j.earscirev.2020.103470

[37] Chen, S., Zhu, Y., Wang, H., Liu, H., Wei, W. and Fang, J. (2011) Shale Gas Reservoir Characterisation: A Typical Case in the Southern Sichuan Basin of China. *Energy*, **36**, 6609-6616. https://doi.org/10.1016/j.energy.2011.09.001

[38] Wang, Y., Liu, L., Zheng, S., Luo, Z., Sheng, Y. and Wang, X. (2019) Full-Scale Pore Structure and Its Controlling Factors of the Wufeng-Longmaxi Shale, Southern Sichuan Basin, China: Implications for Pore Evolution of Highly Overmature Marine Shale. *Journal of Natural Gas Science and Engineering*, **67**, 134-146. https://doi.org/10.1016/j.jngse.2019.04.020

[39] Liang, F., Bai, W., Zou, C., Wang, H., Wu, J., Ma, C., *et al.* (2016) Shale Gas Enrichment Pattern and Exploration Significance of Well Wuxi-2 in Northeast Chongqing, NE Sichuan Basin. *Petroleum Exploration and Development*, **43**, 386-394. https://doi.org/10.1016/s1876-3804(16)30045-3

[40] Dong, D., Liang, F., Guan, Q., Jiang, Y., Zhou, S., Yu, R., *et al.* (2023) Development Model and Identification of Evaluation Technology for Wufeng Formation-Longmaxi Formation Quality Shale Gas Reservoirs in the Sichuan Basin. *Natural Gas Industry B*, **10**, 165-182. https://doi.org/10.1016/j.ngib.2023.02.001

[41] Zhang, C., Dong, D., Wang, Y. and Guan, Q. (2017) Brittleness Evaluation of the Upper Ordovician Wufeng-Lower Silurian Longmaxi Shale in Southern Sichuan Basin, China. *Energy Exploration & Exploitation*, **35**, 430-443. https://doi.org/10.1177/0144598716687929

[42] Li, L., Liu, J., Wang, W., Zhao, P. and Lu, Y. (2023) Variational Reservoir Characteristics of the Wufeng-Longmaxi Formation from Different Sedimentary Region and Its Implications in Southeastern Sichuan Basin, China. *ACS Omega*, **8**, 20684-20696. https://doi.org/10.1021/acsomega.3c01235

[43] Chen, L., Lu, Y., Jiang, S., Li, J., Guo, T. and Luo, C. (2015) Heterogeneity of the Lower Silurian Longmaxi Marine Shale in the Southeast Sichuan Basin of China. *Marine and Petroleum Geology*, **65**, 232-246. https://doi.org/10.1016/j.marpetgeo.2015.04.003

[44] Zhang, L., He, X., Li, X., Li, K., He, J., Zhang, Z., *et al.* (2022) Shale Gas Exploration and Development in the Sichuan Basin: Progress, Challenge and Countermeasures. *Natural Gas Industry B*, **9**, 176-186. https://doi.org/10.1016/j.ngib.2021.08.024

[45] Dong, D., Gao, S., Huang, J., Guan, Q., Wang, S. and Wang, Y. (2015) Discussion on the Exploration & Development Prospect of Shale Gas in the Sichuan Basin. *Natural Gas Industry B*, **2**, 9-23. https://doi.org/10.1016/j.ngib.2015.02.002

[46] Asante-Okyere, S., Marfo, S.A. and Ziggah, Y.Y. (2023) Estimating Total Organic Carbon (TOC) of Shale Rocks from Their Mineral Composition Using Stacking Generalization Approach of Machine Learning. *Upstream Oil and Gas Technology*, **11**, Article ID: 100089. https://doi.org/10.1016/j.upstre.2023.100089

[47] Zhai, G. (1987) The Collection of Chinese Petroleum Geology. Petroleum Industry Press.

[48] Zhen, X., Yu, M., He, X. and Li, S. (2018) Multi-Target Regression via Robust Low-Rank Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 497-504. https://doi.org/10.1109/tpami.2017.2688363

[49] Ali, N.M., Farouk, A.I.B., Haruna, S.I., Alanazi, H., Adamu, M. and Ibrahim, Y.E. (2022) Feature Selection Approach for Failure Mode Detection of Reinforced Concrete Bridge Columns. *Case Studies in Construction Materials*, **17**, e01383.

https://doi.org/10.1016/j.cscm.2022.e01383

[50] Kimura, S. and Tokuhisa, M. (2020) Detection of Weak Relevant Variables Using Random Forests. 2020 59*th Annual Conference of the Society of Instrument and Control Engineers of Japan* (*SICE*), Chiang Mai, 23-26 September 2020, 838-845. https://doi.org/10.23919/sice48898.2020.9240246

[51] She, X., Geng, H., Gou, X. and Chen, K. (2024) A Real-Time Event-Based Learning Approach for Input-Output Sensitivity Analysis. 2024 3*rd Conference on Fully Actuated System Theory and Applications* (*FASTA*), Shenzhen, 10-12 May 2024, 770-774. https://doi.org/10.1109/fasta61401.2024.10595153

[52] Sedgwick, P. (2012) Pearson's Correlation Coefficient. *BMJ*, **345**, e4483. https://doi.org/10.1136/bmj.e4483

[53] Tong, S., Wang, F., Gao, H. and Zhu, W. (2024) A Machine Learning-Based Method for Analyzing Factors Influencing Production Capacity and Production Forecasting in Fractured Tight Oil Reservoirs. *International Journal of Hydrogen Energy*, **70**, 136-145. https://doi.org/10.1016/j.ijhydene.2024.05.036

[54] Li, J.D., *et al.* (2017) Feature Selection: A Data Perspective. *ACM Computing Surveys* (*CSUR*), **50**, 1-45.

[55] Lawal, A., Yang, Y., He, H. and Baisa, N.L. (2024) Machine Learning in Oil and Gas Exploration: A Review. *IEEE Access*, **12**, 19035-19058. https://doi.org/10.1109/access.2023.3349216

[56] Shi, S., Shi, G., Pei, J., Li, L., Zhao, K. and He, Y. (2023) Porosity Prediction in Tight Sandstone Reservoirs Based on a One-Dimensional Convolutional Neural Network-Gated Recurrent Unit Model. *Applied Geophysics*. https://doi.org/10.1007/s11770-023-1044-9

[57] Wang, B., Sharma, J., Chen, J. and Persaud, P. (2021) Ensemble Machine Learning Assisted Reservoir Characterization Using Field Production Data—An Offshore Field Case Study. *Energies*, **14**, Article 1052. https://doi.org/10.3390/en14041052

[58] Tian, F., Fu, Y., Liu, X., Li, D., Jia, Y., Shao, L., *et al.* (2024) A Comprehensive Evaluation of Shale Oil Reservoir Quality. *Processes*, **12**, Article 472. https://doi.org/10.3390/pr12030472

[59] Schober, P., Boer, C. and Schwarte, L.A. (2018) Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, **126**, 1763-1768. https://doi.org/10.1213/ane.0000000000002864

[60] Ali, M., Zhu, P., Huolin, M., Pan, H., Abbas, K., Ashraf, U., *et al.* (2023) A Novel Machine Learning Approach for Detecting Outliers, Rebuilding Well Logs, and Enhancing Reservoir Characterization. *Natural Resources Research*, **32**, 1047-1066. https://doi.org/10.1007/s11053-023-10184-6

[61] Zhen, Y., Zhang, A., Zhao, X., Ge, J., Zhao, Z. and Yang, C. (2024) Prediction of Deep Low Permeability Sandstone Seismic Reservoir Based on CBAM-CNN. *Geoenergy Science and Engineering*, **242**, Article ID: 213241. https://doi.org/10.1016/j.geoen.2024.213241

[62] Saleh, R., Majzoub, S. and Saleh, A.K.M.E. (2025) Fundamentals of Robust Machine Learning: Handling Outliers and Anomalies in Data Science. John Wiley & Sons. https://doi.org/10.1002/9781394294404

[63] Demir, S. and Sahin, E.K. (2023) Application of State-of-the-Art Machine Learning Algorithms for Slope Stability Prediction by Handling Outliers of the Dataset. *Earth Science Informatics*, **16**, 2497-2509. https://doi.org/10.1007/s12145-023-01059-8

[64] Ahmed, H.A., Muhammad Ali, P.J., Faeq, A.K. and Abdullah, S.M. (2022) An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normaliza-

tion Method. *Aro-The Scientific Journal of Koya University*, **10**, 29-37. https://doi.org/10.14500/aro.10970

[65] ur Rehman, A. and Belhaouari, S.B. (2021) Unsupervised Outlier Detection in Multidimensional Data. *Journal of Big Data*, **8**, Article No. 80. https://doi.org/10.1186/s40537-021-00469-z

[66] Fernández, Á., Bella, J. and Dorronsoro, J.R. (2022) Supervised Outlier Detection for Classification and Regression. *Neurocomputing*, **486**, 77-92. https://doi.org/10.1016/j.neucom.2022.02.047

[67] Bhuvana, R., Maheshwari, S. and Sasikala, S. (2024) Outliers-Detection & Implementation Using Machine Learning Algorithm. 2024 13*th International Conference on System Modeling & Advancement in Research Trends* (*SMART*), Moradabad, 6-7 December 2024, 235-238. https://doi.org/10.1109/smart63812.2024.10882474

[68] Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L. and Shao, L. (2023) Normalization Techniques in Training DNNs: Methodology, Analysis and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 10173-10196. https://doi.org/10.1109/tpami.2023.3250241

[69] Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L. and J. Tallón-Ballesteros, A. (2023) The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. In: García Bringas, P., *et al.*, Eds., 18*th International Conference on Soft Computing Models in Industrial and Environmental Applications* (*SOCO* 2023), Springer, 344-353. https://doi.org/10.1007/978-3-031-42536-3_33

[70] Singh, N. and Singh, P. (2021) Exploring the Effect of Normalization on Medical Data Classification. 2021 *International Conference on Artificial Intelligence and Machine Vision* (*AIMV*), Gandhinagar, 24-26 September 2021, 1-5. https://doi.org/10.1109/aimv53313.2021.9670938

[71] Ahmed, H.U., Mostafa, R.R., Mohammed, A., Sihag, P. and Qadir, A. (2022) Support Vector Regression (SVR) and Grey Wolf Optimization (GWO) to Predict the Compressive Strength of GGBFs-Based Geopolymer Concrete. *Neural Computing and Applications*, **35**, 2909-2926. https://doi.org/10.1007/s00521-022-07724-1

[72] Mkono, C.N., Chuanbo, S., Mulashani, A.K., Abelly, E.N., Kasala, E.E., Shanghvi, E.R., *et al.* (2025) Improved Reservoir Porosity Estimation Using an Enhanced Group Method of Data Handling with Differential Evolution Model and Explainable Artificial Intelligence. *SPE Journal*, **30**, 1922-1940. https://doi.org/10.2118/224438-pa

[73] Awad, M. and Khanna, R. (2015) Support Vector Regression. In: Awad, M. and Khanna, R., Eds., *Efficient Learning Machines*, Apress, 67-80. https://doi.org/10.1007/978-1-4302-5990-9_4

[74] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[75] Yu, B., Zhao, H., Tian, J., Liu, C., Song, Z., Liu, Y., *et al.* (2021) Modeling Study of Sandstone Permeability under True Triaxial Stress Based on Backpropagation Neural Network, Genetic Programming, and Multiple Regression Analysis. *Journal of Natural Gas Science and Engineering*, **86**, Article ID: 103742. https://doi.org/10.1016/j.jngse.2020.103742

[76] Hecht-Nielsen, R. (1992) Theory of the Backpropagation Neural Network. In: Wechsler, H., Ed., *Neural Networks for Perception*, Elsevier, 65-93. https://doi.org/10.1016/b978-0-12-741252-8.50010-8

[77] Chen, Y. and Chen, Z. (2022) A Prediction Model of Wall Shear Stress for Ultra-High-Pressure Water-Jet Nozzle Based on Hybrid BP Neural Network. *Engineering*

Applications of Computational Fluid Mechanics, **16**, 1902-1920. https://doi.org/10.1080/19942060.2022.2123404

[78] Song, X., Zhang, Y., Guo, Y., Sun, X. and Wang, Y. (2020) Variable-Size Cooperative Coevolutionary Particle Swarm Optimization for Feature Selection on High-Dimensional Data. *IEEE Transactions on Evolutionary Computation*, **24**, 882-895. https://doi.org/10.1109/tevc.2020.2968743

[79] Majid, A., Mwakipunda, G.C. and Guo, C. (2023) Solution Gas/Oil Ratio Prediction from Pressure/Volume/Temperature Data Using Machine Learning Algorithms. *SPE Journal*, **29**, 999-1014. https://doi.org/10.2118/217979-pa

[80] Frazier, P.I. (2018) A Tutorial on Bayesian Optimization. arXiv: 1807.02811.

[81] Altindag, R. (2009) Assessment of Some Brittleness Indexes in Rock-Drilling Efficiency. *Rock Mechanics and Rock Engineering*, **43**, 361-370. https://doi.org/10.1007/s00603-009-0057-x

[82] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[83] Capitaine, L., Genuer, R. and Thiébaut, R. (2020) Random Forests for High-Dimensional Longitudinal Data. *Statistical Methods in Medical Research*, **30**, 166-184. https://doi.org/10.1177/0962280220946080

[84] Mayapada, R., Susetyo, B. and Sartono, B. (2021) A Comparison between Random Forest and Mixed Effects Random Forest to Predict Students' Math Performance in Indonesia. *International Journal of Sciences: Basic and Applied Research*, **57**, 1-8.

[85] Cochrane, C., Ba, D., Klerman, E.B. and St. Hilaire, M.A. (2021) An Ensemble Mixed Effects Model of Sleep Loss and Performance. *Journal of Theoretical Biology*, **509**, Article ID: 110497. https://doi.org/10.1016/j.jtbi.2020.110497

[86] Mwakipunda, G.C., Ibrahim, A., Kouassi, A.K.F., Komba, N.A., Ayimadu, E.T., Mgimba, M.M., *et al.* (2024) Estimating Carbon Dioxide Solubility in Brine Using Mixed Effects Random Forest Based on Genetic Algorithm: Implications for Carbon Dioxide Sequestration in Saline Aquifers. *SPE Journal*, **29**, 6530-6546. https://doi.org/10.2118/223123-pa

[87] Tong, D., Wang, Y., Ma, J., Wang, J. and Li, J. (2024) Construction of a Prediction Model for Voriconazole-Induced Hepatotoxicity Based on Mixed-Effects Random Forest. In: Craddock, G., Fitzpatrick, D., McHugh, D. and Maher, H., Eds., *Studies in Health Technology and Informatics*, IOS Press, 319-323. https://doi.org/10.3233/shti230979

[88] Katreddi, S., Thiruvengadam, A., Thompson, G.J. and Schmid, N.A. (2023) Mixed Effects Random Forest Model for Maintenance Cost Estimation in Heavy-Duty Vehicles Using Diesel and Alternative Fuels. *IEEE Access*, **11**, 67168-67179. https://doi.org/10.1109/access.2023.3290994

[89] Davoodi, S., Mehrad, M., Wood, D.A., Al-Shargabi, M., Eremyan, G. and Shulgina, T. (2024) A Novel Data-Driven Model for Real-Time Prediction of Static Young's Modulus Applying Mud-Logging Data. *Earth Science Informatics*, **17**, 5771-5793. https://doi.org/10.1007/s12145-024-01474-5

[90] Açıkkar, M. and Altunkol, Y. (2023) A Novel Hybrid PSO- and GS-Based Hyperparameter Optimization Algorithm for Support Vector Regression. *Neural Computing and Applications*, **35**, 19961-19977. https://doi.org/10.1007/s00521-023-08805-5

[91] Cho, H., Kim, Y., Lee, E., Choi, D., Lee, Y. and Rhee, W. (2020) Basic Enhancement Strategies When Using Bayesian Optimization for Hyperparameter Tuning of Deep Neural Networks. *IEEE Access*, **8**, 52588-52608.

https://doi.org/10.1109/access.2020.2981072

[92] Egelé, R., Guyon, I., Vishwanath, V. and Balaprakash, P. (2023) Asynchronous Decentralized Bayesian Optimization for Large Scale Hyperparameter Optimization. 2023 *IEEE* 19*th International Conference on e-Science* (*e-Science*), Limassol, 9-13 October 2023, 1-10. https://doi.org/10.1109/e-science58273.2023.10254839

[93] Li, K., Yin, Z. and Liu, Y. (2024) A Hybrid SVR-BO Model for Predicting the Soil Thermal Conductivity with Uncertainty. *Canadian Geotechnical Journal*, **61**, 258-274. https://doi.org/10.1139/cgj-2023-0105

[94] Ramkumar, D. and Jothiprakash, V. (2024) Forecasting Influent Wastewater Quality by Chaos Coupled Machine Learning Optimized with Bayesian Algorithm. *Journal of Water Process Engineering*, **61**, Article ID: 105306. https://doi.org/10.1016/j.jwpe.2024.105306

[95] Wadood, A., Albalawi, H., Alatwi, A.M., Anwar, H. and Ali, T. (2025) Design of a Novel Fractional Whale Optimization-Enhanced Support Vector Regression (FWOA-SVR) Model for Accurate Solar Energy Forecasting. *Fractal and Fractional*, **9**, Article 35. https://doi.org/10.3390/fractalfract9010035

[96] Lorentzen, M., Bredesen, K., Mosegaard, K. and Nielsen, L. (2022) Estimation of Shear Sonic Logs in the Heterogeneous and Fractured Lower Cretaceous of the Danish North Sea Using Supervised Learning. *Geophysical Prospecting*, **70**, 1410-1431. https://doi.org/10.1111/1365-2478.13252

[97] Yadav, S. and Shukla, S. (2016) Analysis of K-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. 2016 *IEEE* 6*th International Conference on Advanced Computing* (*IACC*), Bhimavaram, 27-28 February 2016, 78-83. https://doi.org/10.1109/iacc.2016.25

[98] Yan, T., Shen, S., Zhou, A. and Chen, X. (2022) Prediction of Geological Characteristics from Shield Operational Parameters by Integrating Grid Search and K-Fold Cross Validation into Stacking Classification Algorithm. *Journal of Rock Mechanics and Geotechnical Engineering*, **14**, 1292-1303. https://doi.org/10.1016/j.jrmge.2022.03.002

[99] Zhu, R., Guo, W. and Gong, X. (2019) Short-Term Photovoltaic Power Output Prediction Based on K-Fold Cross-Validation and an Ensemble Model. *Energies*, **12**, Article 1220. https://doi.org/10.3390/en12071220

[100] Vu, H.L., Ng, K.T.W., Richter, A. and An, C. (2022) Analysis of Input Set Characteristics and Variances on K-Fold Cross Validation for a Recurrent Neural Network Model on Waste Disposal Rate Estimation. *Journal of Environmental Management*, **311**, Article ID: 114869. https://doi.org/10.1016/j.jenvman.2022.114869

[101] Andrew, E., Xu, C., Odi, U., Sheludko, S., Silver, A. and Zayer, Y. (2024) Machine Learning-Assisted Petrophysical Rock Type Classification and Permeability Estimation with Flow Zone Indicators. *International Petroleum Technology Conference*, Dhahran, 12 February 2024, IPTC-23537-EA. https://doi.org/10.2523/iptc-23537-ea

[102] Lumumba, V., Kiprotich, D., Mpaine, M., Makena, N. and Kavita, M. (2024) Comparative Analysis of Cross-Validation Techniques: LOOCV, K-Folds Cross-Validation, and Repeated K-Folds Cross-Validation in Machine Learning Models. *American Journal of Theoretical and Applied Statistics*, **13**, 127-137. https://doi.org/10.11648/j.ajtas.20241305.13

[103] Jiang, G. and Wang, W. (2017) Error Estimation Based on Variance Analysis of K-Fold Cross-Validation. *Pattern Recognition*, **69**, 94-106. https://doi.org/10.1016/j.patcog.2017.03.025

[104] Marcot, B.G. and Hanea, A.M. (2020) What Is an Optimal Value of K in K-Fold

Cross-Validation in Discrete Bayesian Network Analysis? *Computational Statistics*, **36**, 2009-2031. https://doi.org/10.1007/s00180-020-00999-9

[105] Jung, Y. and Hu, J. (2015) A *K*-Fold Averaging Cross-Validation Procedure. *Journal of Nonparametric Statistics*, **27**, 167-179. https://doi.org/10.1080/10485252.2015.1010532

[106] Pohjankukka, J., Pahikkala, T., Nevalainen, P. and Heikkonen, J. (2017) Estimating the Prediction Performance of Spatial Models via Spatial K-Fold Cross Validation. *International Journal of Geographical Information Science*, **31**, 2001-2019. https://doi.org/10.1080/13658816.2017.1346255

[107] Fushiki, T. (2009) Estimation of Prediction Error by Using K-Fold Cross-Validation. *Statistics and Computing*, **21**, 137-146. https://doi.org/10.1007/s11222-009-9153-8

[108] Taylor, K.E. (2001) Summarizing Multiple Aspects of Model Performance in a Single Diagram. *Journal of Geophysical Research*: *Atmospheres*, **106**, 7183-7192. https://doi.org/10.1029/2000jd900719

[109] Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G.A. (2010) Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, **23**, 2739-2758. https://doi.org/10.1175/2009jcli3361.1

[110] Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F. (2009) Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling. *Journal of Hydrology*, **377**, 80-91. https://doi.org/10.1016/j.jhydrol.2009.08.003

[111] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. arXiv: 1705.07874.

[112] Mkono, C.N., Shen, C., Mulashani, A.K., Ngata, M.R. and Hussain, W. (2024) A Novel Hybrid Machine Learning Approach and Basin Modeling for Thermal Maturity Estimation of Source Rocks in Mandawa Basin, East Africa. *Natural Resources Research*, **33**, 2089-2112. https://doi.org/10.1007/s11053-024-10372-y

[113] Zhao, X., Chen, X., Huang, Q., Lan, Z., Wang, X. and Yao, G. (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**, Article ID: 110517. https://doi.org/10.1016/j.petrol.2022.110517

[114] Molnar, C. (2020) Interpretable Machine Learning. Lulu.com.

[115] Ye, Y., Tang, S., Xi, Z., Jiang, D. and Duan, Y. (2022) Quartz Types in the Wufeng-Longmaxi Formations in Southern China: Implications for Porosity Evolution and Shale Brittleness. *Marine and Petroleum Geology*, **137**, Article ID: 105479. https://doi.org/10.1016/j.marpetgeo.2021.105479

[116] Zhang, K., *et al.* (2023) Quartz Types, Genesis and Their Geological Significance within the Wufeng-Longmaxi Formation in North-Western Hunan, China. *Frontiers of Earth Science*, **17**, 455-469.

[117] Jia, Y., Tang, J., Lu, Y. and Lu, Z. (2021) Laboratory Geomechanical and Petrophysical Characterization of Longmaxi Shale Properties in Lower Silurian Formation, China. *Marine and Petroleum Geology*, **124**, Article ID: 104800. https://doi.org/10.1016/j.marpetgeo.2020.104800

[118] Guan, Q., Dong, D., Zhang, H., Sun, S., Zhang, S. and Guo, W. (2021) Types of Biogenic Quartz and Its Coupling Storage Mechanism in Organic-Rich Shales: A Case Study of the Upper Ordovician Wufeng Formation to Lower Silurian Longmaxi Formation in the Sichuan Basin, SW China. *Petroleum Exploration and Development*,

48, 813-823. https://doi.org/10.1016/s1876-3804(21)60068-x

[119] Feng, Y., Xiao, X., Wang, E., Hu, D., Liu, R. and Zhou, Q. (2024) Quartz Crystallinity Characteristics and Their Effects on Shale Gas Reservoir Performance: A Case Study of the Deep Longmaxi Formation Shale in the Sichuan Basin, China. *ACS Earth and Space Chemistry*, **8**, 1324-1338. https://doi.org/10.1021/acsearthspacechem.4c00025

[120] Tuzingila, R.M., Kong, L. and Kasongo, R.K. (2025) Experimental Techniques and Rock Physics Modeling to Study the Effects of Mineral Content and Thermal Maturation on Elastic Properties of the Organic-Rich Shale: Case Study of the Longmaxi Formation. *SPE Journal*, **30**, 1884-1905. https://doi.org/10.2118/224419-pa

[121] Lin, J., Hu, H., Zhang, L., Zhang, X. and Liu, Z. (2022) The Maturity of Silurian Longmaxi Shale in Jiaoshiba, Sichuan Basin: As Revealed by Laser Raman Spectroscopy. *Journal of Petroleum Exploration and Production Technology*, **12**, 2609-2624. https://doi.org/10.1007/s13202-022-01516-y

[122] Shi, X., Wu, W., Xu, L., Yin, Y., Yang, Y., Liu, J., *et al.* (2024) Thermal Maturity Constraint Effect and Development Model of Shale Pore Structure: A Case Study of Longmaxi Formation Shale in Southern Sichuan Basin, China. *Minerals*, **14**, Article 163. https://doi.org/10.3390/min14020163

[123] Jamaluddin, Ryka, H. and Hamsir, (2020). Estimating Shale Brittleness Index Based on Mineral Content Using Wireline Log. 3*rd Asia Pacific Meeting on near Surface Geoscience & Engineering*, Chiang Mai, 20-22 April 2020, 1-5. https://doi.org/10.3997/2214-4609.202071028

[124] Wood, D.A. (2023) Geomechanical Brittleness Index Prediction for the Marcellus Shale Exploiting Well-Log Attributes. *Results in Engineering*, **17**, Article ID: 100846. https://doi.org/10.1016/j.rineng.2022.100846

[125] Barnhoorn, A., Primarini, M. and Houben. M. (2016) Fracturing and Brittleness Index Analyses of Shales. *Geophysical Research Abstracts*, **18**, EGU2016-5869.

[126] Mustafa, A., Tariq, Z., Abdulraheem, A., Mahmoud, M., Kalam, S. and Khan, R.A. (2022) Shale Brittleness Prediction Using Machine Learning—A Middle East Basin Case Study. *AAPG Bulletin*, **106**, 2275-2296. https://doi.org/10.1306/12162120181

List of Symbols

IQR	Interquartile range
Q1	25th percentile, the lower quartile
Q3	75th percentile, the upper quartile
DTC	Compressional wave slowness
DTS	Shear sonic log
K	Bulk modulus
E	Young's modulus
v	Poisson ratio
PSO	Particle swarm optimization
MERF	Mixed effect random forest
BO	Bayesian optimization
RMSE	Root mean square error
R^2	Coefficient of determination
MAE	Mean absolute error
SHAP	Shapley additive explanation
σ_i	Measured brittleness index value
$\bar{\sigma}$	Average value of the measured brittleness index
N	Total number of data points
σ_i'	Predicted brittleness index value
$\overline{\sigma'}$	Average value of predicted brittleness index