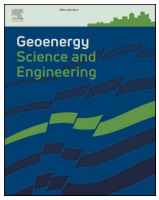
233 (2024) 212518

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

journal homepage: www.sciencedirect.com/journal/geoenergy-science-and-engineering



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

Mbula Ngoy Nadege [a,b], Shu Jiang [a,*], Grant Charles Mwakipunda [a], Allou Koffi Franck Kouassi [a], Paulin Kavuba Harold [a,b], Konan Yao Hugues Roland [a]

[a] Key Laboratory of Theory and Technology of Petroleum Exploration and Development in Hubei Province, China University of Geosciences, Wuhan, 430074, China
[b] Department of Exploration and Production, Faculty of Oil,Gas and New Energies, University of Kinshasa, B. P. 127 Kinshasa XI, Congo

ARTICLE INFO

Keywords:
Brittleness index
Fractures
Shale gas
Random forest based on particle swarm optimization

ABSTRACT

The right placement of fractures helps to enhance gas production in shale gas reservoirs. One parameter that helps to determine the target layers to place hydraulic fractures is the Brittlenex index (BI). However, no universal and appropriate methods can be used to compute BI, with all established correlations being used under different conditions. This paper uses machine learning (ML) methods to predict the BI of Upper Ordovician Wufeng to Lower Silurian Longmaxi formation in the Weiyuan shale gas field, Sichuan Basin, China. Random forest based on particle swarm optimization (PSO-RF) was utilized for the first time to predict BI due to its ability to capture nonlinear relationships between many variables in the dataset, thus giving more accurate results than other models. Collected secondary data from the WY1 well were used for training, whereas WY2 well data were used for testing. The results revealed that PSO-RF outperformed Extreme gradient boosting (XGBoost), Light gradient boosting machine (LightGBM), and K-nearest neighbor (KNN) in predicting BI with high accuracy and minimum errors during training and testing. PSO-RF coefficient of determination (R^2), root mean square error (RMSE), and mean absolute errors (MAE) after training and testing were 0.9934 and 0.9533,4.6327 and 15.5308,2.0974 and 5.3896, respectively. In addition, the best-developed PSO-RF model was used to predict BI in WY3 and WY4 wells for model results validation; it was found that the model predicted the BI with high accuracy. This confirms that the developed model can be used to predict the BI of new development wells without depending on laboratory measurements, which are expensive and time-consuming to compute; thus, the developed model can be adopted as an alternative technique to determine the sweet spot for hydraulic fracturing in shale gas reservoirs to enhance gas production.

1. Introduction

Recently, global energy demand has increased faster because of population growth and rising prosperity led by Asian developing countries (Guan et al., 2023; Shalaeva et al., 2020). In 2021, the US Energy Information Administration (EIA) predicted that the energy demand will increase by 47% in 2050, with oil remaining the top source over renewables (Meghan and Maya, 2021). Conventional reservoirs are the major sources of energy and are depleting fast. However, due to technological developments such as horizontal drilling and hydraulic fracturing, unconventional resources have emerged among the new energy sources. Oil and gas widely use this technology to optimize hydrocarbon production by generating intricate and effective fracture networks to ease fluid flows (Merzoug and Ellafi, 2023; Qun et al., 2022). Previous literature has revealed that the success of hydraulic fracturing depends on rock brittleness (Peng et al., 2022; Shi et al., 2021). Highly brittle shales are thought to benefit from fracturing stimulation (Ore and Gao, 2023; Ye et al., 2022). Researchers have done several investigations regarding the brittleness index (BI) notion. However, it is important to note that a universally accepted definition

* Corresponding author.
E-mail address: jiangsu@cug.edu.cn (S. Jiang).

https://doi.org/10.1016/j.geoen.2023.212518
Received 2 October 2023; Received in revised form 1 November 2023; Accepted 17 November 2023
Available online 25 November 2023

and standardized method for evaluating BI have not yet been established due to diverse physical factors (Meng et al., 2021a; Shi et al., 2017).

There are several established categories of BI expressions based on different conditions, such as mineral rock constituents (Jarvie et al., 2007; Kivi et al., 2018; Kuang et al., 2021; Meng et al., 2021a; Munoz et al., 2016; Rybacki et al., 2015, 2016; Song et al., 2023), based on rock strength properties (Hucka and Das, 1974; Kivi et al., 2018; Kuang et al., 2021; Li, 2022; Li et al., 2017; Xie et al., 2022; Zhang et al., 2016), based on rock stress-strain response to deviatoric loading elastic properties (Gao et al., 2023; Kivi et al., 2018; Kuang et al., 2021; Peng et al., 2022; Rickman et al., 2008; Rybacki et al., 2016), stress-strain characteristics (Andreev, 1995; Khan et al., 2023; Kivi et al., 2018; Kuang et al., 2021; Nouri et al., 2022; Xia et al., 2022), and energy balance analysis (Gong and Wang, 2022; Khan et al., 2023; Kivi et al., 2018; Kuang et al., 2021; Xia et al., 2017; Xu et al., 2022) as shown in Table 1.

Due to the poor performance of developed empirical correlations, machine learning (ML) has been used recently to forecast the BI of rock formations (Cornelio and Ershaghi, 2019; Hassan et al., 2022; Mustafa et al., 2022; Ore and Gao, 2023; Shi et al., 2016b; Sun et al., 2020). For instance, Ye et al. (2022) used a backpropagation neural network (BPNN) integrated with principal component analysis (PCA) to predict experimental BI from well logs data collected from the Wufeng-Longmaxi and Baota formations in the Sichuan Basin. PCA-BPNN was compared with BPNN. The well-log data used in their study includes density (DEN), neutron porosity (CNL), gamma ray (GR), spontaneous potential (SP), compressional wave slowness (DT), and deep laterolog (LLD). The results revealed that BI predicted by PCA-BPNN matched the experimental data compared to BPNN because PCA eliminated LLD as input during training due to its small correlation with BI. Also, Shi et al. (2017) predicted BI from conventional well logs and petrophysical data collected from tight oil formation in the Xinjiang Basin, China, using multilayer perception neural network (MLPNN) and radial basis function neural network (RBFNN) techniques. The inputs data used for the models' training and testing were GR, LLD, shallow laterolog (LLS), compessated acoustic (AC), lithology density (DEN), neutron porosity (CNL), and natural spectrum logs including Uranium (U), Thorium (Th), and Potassium (K). The results revealed that RBFNN outperformed MLPNN with small errors and high correlation coefficient. Further, Zhang et al. (2022) utilized hybrid machine learning, i.e., sparrow search algorithm-extreme learning machine (SSA-ELM) in predicting BI from well logs data collected from Songliao Basin found in northeast China in which other wells were used for training the model with two wells utilized for testing the model. GR, DEN, AC, CNL, RLLD and RLLS well logs were available data in which GR was excluded during training due to a small correlation with BI. It was found that hybrid SSA-ELM outperformed other used models' such as BPNN, ELM, particle swarm optimization-ELM (PSO-ELM), support vector machine (SVM), conventional neural network (CNN), random forest (RF), long short-term memory (LSTM), K-nearest neighbor (KNN), decision tree (DT), kernel-based ELM (KELM), and grey wolf optimizer-ELM (GWO-ELM). Furthermore, Lee and Lumley (2023) predicted shale mineralogical BI (MBI) from seismic and elastic property well logs of 13 wells in the Wolfcamp shale of the Midland Basin, West Texas, using DT, SVM, ensembles, multiple linear regression (MLR), and neural network. The inputs for the models' were GR, DEN, AC, P-wave velocity (V_P), conductivity, S-wave velocity (V_S), U, bulk density (ρ), poisson ratio (v), V_P/V_S ratio, and Young's modulus (E). It was revealed that MLR outperformed other models', with E and V_P having a high correlation with MBI while V_P/V_S ratio and v had a low correlation. Also, Shi et al. (2016a) predicted rock BI of Silurian Longmaxi black shales of well J1 from the Jiaoshiba area of SE Sichuan Basin, south China, using BPNN, ELM, and regression models'. The inputs for the models' were DEN, DTC, spontaneous potential (SP), LLD, and CNL. It was found that BPNN outperformed other models' by giving low errors with high correlation coefficient while ELM took a short running time. In addition, Kivi et al. (2017) developed an adaptive neuro-fuzzy inference system (ANFIS) to

Table 1
Outlines of often utilized empirical brittleness indices (BI) correlations.

Measurement's methods	Expressions	Symbols definitions
Based on rock mineral constituents	$BI_1 = \frac{W_{qtz}}{W_{qtz+carb+cly}}$ $BI_2 = \frac{W_{qtz+dol}}{W_{qtz+carb+cly+TOC}}$ $BI_3 = \frac{W_{QFM+carb}}{W_{total}}$ $BI_4 = \frac{a_{QFP}V_{QFP}}{a_{QFP}V_{QFP} + a_{carb}V_{carb} + a_{cly+TOC}V_{cly+TOC} + a_{\varphi}\varphi}$	W_x is weight fraction of component x V_x is volume fraction of component x a_x is weighting factor of component x Q_{tz} is quartz, carb is carbonate, cly is clay Dol is dolomite, TOC is total organic carbon QFM are quartz, feldspar, and mica QFP are quartz, feldspar, and pyrite φ is porosity T_0 is unconfined tensile strength UCS is unconfined compressive strength φ is internal friction angle E is Young's modulus v is Poisson's ratio
Based on rock strength properties	$BI_5 = \frac{UCS - T_0}{UCS + T_0}$ $BI_6 = \frac{UCS}{T_0}$ $BI_7 = \frac{UCS.T_0}{2}$ $BI_8 = \sin(\varphi)$	
Based on rock stress-strain response to deviatoric loading, elastic properties	$BI_9 = \frac{1}{2}\left(\frac{E - E_{min}}{E_{max} - E_{min}} + \frac{\nu_{max} - \nu}{\nu_{max} - \nu_{min}}\right)$	E_{min} and E_{max} are minimum and maximum Young's modulus
Stress-strain characteristics	$BI_{10} = \varepsilon_p$ $BI_{11} = \frac{\varepsilon_e}{\varepsilon_f}$ $BI_{12} = \frac{\sigma_f - \sigma_r}{\sigma_f}$ $BI_{13} = \frac{\varepsilon_r - \varepsilon_f}{\varepsilon_r}$ $BI_{14} = \frac{\varepsilon_f^p - \varepsilon_c^p}{\varepsilon_c^p}$ $BI_{15} = \frac{H}{E}$ $BI_{16} = \frac{\sigma_f - \sigma_r}{\sigma_f}\frac{1}{10}\log\left\|\frac{\sigma_{cd} - \sigma_r}{\varepsilon_{cd} - \varepsilon_r}\right\|$	ε_p is sustained plastic strain at failure ε_e is total elastic at failure ε_f is total strain at failure σ_f is stress at failure σ_r is residual strength ε_r is residual strain ε_f^p is plastic strain when frictional strength is fully mobilized ε_c^p is plastic strain when cohesive

(continued on next page)

Table 1 (*continued*)

Measurement's methods	Expressions	Symbols definitions
Energy balance analysis	$BI_{17} = \frac{dW_{et}}{dW_{et} + dW_p}$ $BI_{18} = \frac{dW_r}{dW_e} = \frac{M - E}{M}$ $BI_{19} = \frac{dW_a}{dW_e}$ $BI_{20} = \frac{dW_{et}}{dW_p + dW_r}$ $BI_{21} = \frac{dW_{et} + dW_p}{dW_p + dW_r}$ $BI_{22} = \frac{dW_{et}}{dW_r}$ $BI_{23} = \frac{dW_p + dW_r}{dW_e + dW_p}$ $BI_{24} = \frac{dW_a}{dW_e + dW_p}$ $BI_{25} = BI_E + BI_{post} = \frac{(\sigma_f - \sigma_r)(\varepsilon_r - \varepsilon_f)}{\sigma_f \varepsilon_f} + \frac{\sigma_f - \sigma_r}{\varepsilon_r - \varepsilon_f}$ $BI_{new} = \frac{1}{2}(BI_{new-1} + BI_{new-II}) = \frac{1}{2}\left(\frac{dW_e}{dW_r} + \frac{dW_e}{dW_{et} + dW_p}\right)$	strength is completely degraded H is hardening modulus σ_{cd} is crack damage stress or yield stress ε_{cd} is crack damage strain or yield strain dW_{et} is total elastic energy dW_p is plastic energy M is post-peak modulus dW_r is rupture energy dW_e is consumed elastic energy dW_a is additional energy

predict BI from well-log data collected from a western Iranian Basin. The inputs used were (GR), density (RHOB), neutron porosity (NPHI), slowness of the compressional wave (DTC), and electrical resistivity (RT). It was revealed that ANFIS model outperformed empirical correlations (BI_1 to BI_9). Further, it was found that BI has positive correlation only with RHOB logs whilst others had negative correlations. In addition, Table 2 summarize few previous studies on predicting BI for shale formations which are essential in locating the favourable place for hydraulic fracturing. However, these ML models' have several limitations such as overfitting, poor prediction of outputs which do not match the original data, computational complexity, bias and fairness, interpretability etc.

Hence, this paper utilizes ML techniques to locate where hydraulic fracturing can be executed to enhance hydrocarbon production from shale gas formations in Weiyuan gas fields, Sichuan Basin, China, by predicting BI. To the best of the author's knowledge, particle swarm optimization-random forest (PSO-RF) was used for the first time to predict the BI of the shale formations in which higher BI formations are preferred for hydraulic fracturing operations because it indicates that the rock formation has properties that make it conducive to fracturing and creating effective pathways for the flow of reservoir fluids. However, it is important to note that hydraulic fracturing is a complex process, and the suitability of a rock formation for fracking is influenced by several factors, including not only the brittle index but also the depth, thickness, and composition of the rock, as well as the presence of natural fractures and faults. To assess PSO-RF effectiveness in predicting BI, it was compared with Extreme gradient boosting (XGBoost), Light gradient boosting machine (LightGBM), and K-nearest neighbor (KNN). The results of this paper helped to develop an appropriate ML technique that helped to predict the BI of shale formations that will help to increase hydraulic fracturing effectiveness if properly located, which will enhance shale gas production. This paper contains several sections: introduction, geological setting, data preprocessing,machine learning algorithms, results and discussions, and conclusions.

2. Geological setting

The study area of this paper is the Weiyuan shale gas field located in the Sichuan Basin, as shown in Fig. 1. The Sichuan Basin is found in the southwestern region of China. The Basin is located in the northwest and is part of the Yangtze Platform. It is a large intracratonic Basin on the stable South China block. It is located in Sichuan province and the Chongqing area and encompasses an area of approximately 23×10^4 km^2 (Mgimba et al., 2022). The Longmaxi and Wufeng shales are the principal focal points for the exploration and development of shale gas within the southeastern region of Chongqing. These formations exhibit abundant organic material and undergo substantial thermal modification at a considerable depth. The primary factors that contributed to these formations' significant shale gas formation were supported by the persistent anaerobic conditions that prevailed over an extended period (Wang et al., 2021). The Sichuan Basin today generates the most gas in China and has the country's most abundant natural gas resources. In 2014, it was reported that the Sichuan Basin has 3.22×10^{12} m^3 gas reserves (Mgimba et al., 2023a). The Basin formed in the late Proterozoic and has continued to the present day. The basement near the Basin's centre comprises extensively metamorphosed intermediate basalt magmatic rocks. The Sichuan Basin is surrounded by mountains such as Wu on the eastern and Daha on the northern part. The Micang and Daba, Daliang, Longmen, and Dalou mountains border the Basin to the north, south, west, and east (Mgimba et al., 2023b).

The Weiyuan region is found in the southwestern part of the Sichuan Basin. It includes Weiyuan County, Gongxian County, and Rongxian County, all located inside Neijiang City. The site is located inside the geologically elevated region of central Sichuan, characterised by relatively low levels of structural strength. The fundamental framework of the gas field exhibits characteristics of a broad and gradual anticline formation. The burial depth of the Wufeng-Longmaxi formation bottom ranges from 1100 to 2800 m, with a steady increase from the northwest to the southeast (Chen et al., 2019b; Zeng et al., 2018). The Weiyuan region is tectonically found in the southwestern section of the central plain zone of the paleouplift in central Sichuan. It consists mostly of the slope area on the eastern wing of the Weiyuan anticline (Huang et al., 2012; Meng et al., 2021b). Several formations were created in the southeast of Chongqing during the Cambrian, Ordovician, and Silurian epochs, while others are absent owing to tectonic uplift and erosion. As shown in Fig. 2, the principal source rocks in the area are the Lower Cambrian Niutitang formation, Lower Silurian Longmaxi formation, and Upper Ordovician Wufeng formation (Mgimba et al., 2023a).

3. Data preprocessing

3.1. Data source and preprocessing

The study area of this paper includes four wells in which WY1 was used for training, WY2 was used for testing the models', and two other wells (WY3 and WY4) were used for model validation of the best-developed model. These wells located in the Weiyuan shale gas field were used to determine the BI of the formation, which is one of the important parameters to locate the right layers for hydraulic fracturing operations to enhance shale gas production. The inputs of the models' include young modulus (E), bulk modulus (K), shear modulus (G), compressional wave slowness (DTC), shear sonic log (DTS), resistivity log (R), poisson ratio (v) whilst the output of the model was brittleness index (BI). The statistical analysis of the datasets used for training and testing are shown in Tables 3 and 4. To improve the models' accuracy and robustness, avoiding overfitting and biasness, a modified Z-scores method was used to determine outliers of the data as presented in box

Table 2
Summary of a few different ML models' used to predict BI of shale formations.

References	Inputs	Methods	Remarks	Limitations of the best method
Sun et al. (2020)	Density, Schmidt hammer rebound number, point load index and p-wave velocity	Chi-square automatic interaction detector (CHAID), RF, SVM, KNN, and ANN.	-RF outperformed other methods in accuracy, followed by ANN and KNN.	-Overfitting when the number of trees are large -Sensitivity to noise data - Favours the majority class, hindering minority performance.
Mustafa et al. (2022)	GR, DTC, resistivity log, RHOB, and NPHI.	Feed forward ANN (FFANN) and ANFIS.	-Both ANFIS and FFANN can be utilized for BI prediction. -ANFIS performed better than FFANN due to the least errors and high correlation during the training and testing of the models'.	- ANFIS models' may exhibit a high level of complexity and pose challenges in terms of interpretability, particularly in scenarios where the input data comprises many fuzzy rules and multiple layers of adaptive nodes. - ANFIS is prone to overfitting, particularly when there is little control over the complexity of the model. - ANFIS models' have several hyperparameters that need to be tuned carefully, such as the number of fuzzy rules, the types of membership functions, and the learning rates. Choosing the right hyperparameters can be hard and greatly affect how well the model performs.
Ore and Gao (2023)	GR, DTC, RHOB, caliper (HCAL), NPHI, and photoelectric factor (PEZ)	Gradient boosting (XGBoost), support vector regression (SVR), and neural networks (NN)	XGBoost outperformed other models', followed by NN and SVR for training and testing.	- Computational complexity - Overfitting problem -Hyperparameter sensitivity
Shi et al. (2016b)	GR, CNL, LLD, U, DTC, spontaneous potential (SP)	Back propagation artificial neural network (BPANN) and least squares support vector regression (LS-SVR)	-LS-SVR outperformed BPANN but is time-consuming because it needs parameter determination for the validation process. -GR and U have the best correlation with BI, whereas LLD has a weak correlation with BI.	- LS-SVR detects outliers. Outliers can affect model training and produce unworthy outcomes. LS-SVR employs kernel functions to transform input data into higher-dimensional feature spaces. Selecting an appropriate kernel function can be difficult, requiring domain knowledge and experience. - Since LS-SVR does not natively produce probabilistic outputs, estimating prediction uncertainty can be difficult.
Hassan et al. (2022)	Mineralogical composition Ca, Na, Si, Al, and K	ANN, fuzzy logic (FL), and SVM	-All the developed models' can be used for MBI prediction. -ANN outperformed all the models', followed by SVM, then FL.	-Require large datasets to provide accurate results - ANNs have a lot of hyperparameters that must be carefully chosen, such as the number of layers, neurons per layer, learning rate, and so on. Finding the best set of hyperparameters can be a time-consuming and expensive procedure. -Overfitting problem
Cornelio and Ershaghi (2019)	GR, HCAL, DTC, DTS, RHOB, RLLD, RLLS, and medium resistivity.	KNN, SVR, DT, RF, and GB regressions.	KNN outperformed other models'	- KNN has a significant computational cost during training and testing when working with huge datasets. - KNN memorizes the complete dataset rather than learning a discriminative function. As a result, it may have difficulty generalizing to new data, resulting in overfitting the training set.

plots in Fig. 3 (data with outliers) and Fig. 4 (data without outliers) (Jamshidi et al., 2022; Sarvi et al., 2022; Yaro et al., 2023). Modified Z-score formula incorporates the median and median absolute deviation in a robust Z-score formula against outliers. This method assumes that the data are normally distributed. Furthermore, to ensure that the ML model treats and handles the inputs and output data impartially and avoids bias and overfitting, all datasets were standardized using the min-max approach to values between 0 and 1 using Eq. (1)(Majid et al., 2023; Mulashani et al., 2022).

$$y_i^{'} = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \tag{1}$$

Where $y_i^{'}$, y_i, $y_{\min}$, $y_{\max}$ are the normalized value of y_i, the value to be normalized, the minimum value of y_i, and the maximum value of y_i, respectively.

3.2. Inputs-output correlations

Correlation is a statistical measure describing the degree to which two variables change together (Niaki et al., 2023). The correlation between the inputs and output can be positive, negative, or zero relationships. Positive correlations occur when an increase in one variable is associated with an increase in another variable. In comparison, a negative correlation occurs when an increase in one variable is associated with a decrease in another variable. Zero correlation occurs when changes in the input variable do not consistently predict changes in the output variable. This section analyzed linear relationships between inputs and output in cross plots (Dev et al., 2022; Mangalathu et al., 2022; Ryu et al., 2022). From Fig. 5, it has been shown that young modulus (E), bulk modulus (K), shear modulus (G), resistivity log (R), and poisson ratio (v) have a positive relationship with BI. In contrast, compressional wave slowness (DTC) and shear sonic log (DTS) negatively correlate with BI. The quantitative correlation between the datasets is shown in Fig. 6. Parameters with higher absolute magnitudes of relevancy factors demonstrate greater importance in estimating the BI(Ye et al., 2022). In addition, young modulus (E) and shear modulus (G) have more influence on BI, with a correlation coefficient of 0.8947 and 0.8747, respectively. In contrast, resistivity log (R) and poison ratio (v) have a weak influence on BI, with a correlation coefficient of 0.2757 and 0.1974, respectively. Despite resistivity log (R) and poisson ratio (v) having a weak influence on BI, they were included during model training with other inputs because they have a correlation coefficient of greater than 0.1 (Ore and Gao, 2023).

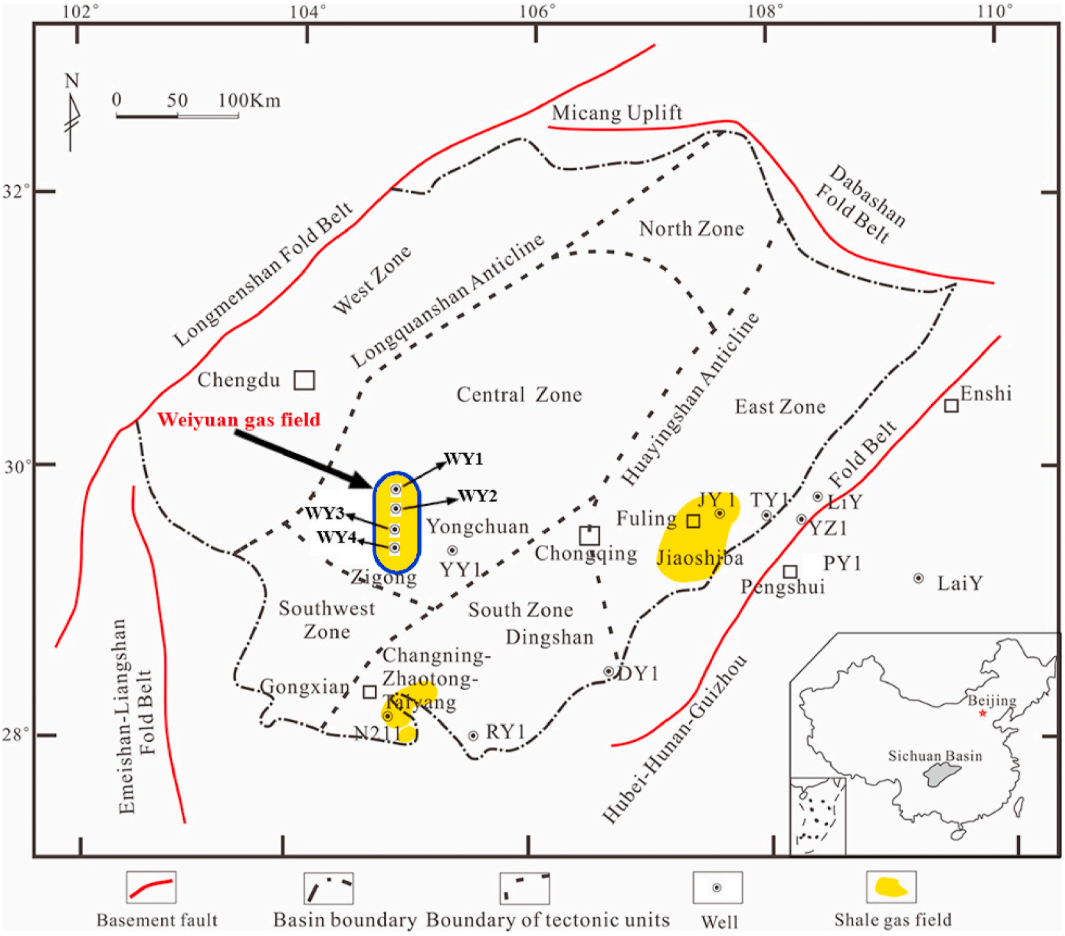


Fig. 1. Locations of Weiyuan Shale Gas Field and shale gas wells (Nie et al., 2021).

4. Machine learning algorithms

This section of the paper discusses four different ML models' to assess their effectiveness in predicting the BI of Weiyuan shale gas fields. After training and testing the datasets, the models' were compared, and the best model was used to predict BI in the other two pilot test wells.

4.1. Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a popular open-source Gradient Boosting Decision Tree (GBRT) widely used in machine learning competitions and real-world applications. Like LightGBM, XGBoost is designed to optimize and enhance the performance of gradient-boosting algorithms. It is known for its efficiency, scalability, and versatility. It applies to both regression and classification problems. It has gained popularity due to its ability to deliver strong predictive performance with relatively little parameter tuning and its compatibility with various platforms and languages. The ensemble technique leverages the capabilities of weak learners to attain robust performance. It is characterised by its high speed and efficiency in mitigating overfitting. Additionally, the model incorporates a novel tree model that allows users to build their loss function. Column sub-sampling and shrinking techniques are used to reduce both the variance and bias of the model (Chen et al., 2019a; Chen and Guestrin, 2016; Kavzoglu and Teke, 2022; Liu et al., 2022; Osman et al., 2021). Suppose we have a data set (DS) with n samples and m features $DS = \{(x_i, y_i : i = 1,\ldots,n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})\}$. Let $\hat{y}_i$ be the predicted output of the model defined as (Alabdullah et al., 2022; Liu et al., 2023):

$$\hat{y}_i = \sum_{i=1}^{K} f_k(x_i), f_k \in F \tag{2}$$

Where K, x_i, and F are the number of regression trees, features related to sample i, and the space of regression trees, respectively. f_k is the weight of the leaf for node j. The objective function of XGBoost, which needs to be minimized, is defined as (Alabdullah et al., 2022; Liu et al., 2023):

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{4}$$

Where $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ represents the training loss function, γ represents the degree of regularization, K is the number of trees, λ represents the regularization coefficient, ω stands for leaf weight, and $\Omega(f)$ is a parameter used to limit the complexity of the model and prevent the model from overfitting. To minimize the objective function, assume, $\hat{y}_i^t$,*i-th* output of the model at the t-iteration. To minimize the following objective, the branches, f_t, need to be added (Alabdullah et al., 2022; Liu et al., 2023).

$$Obj^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{5}$$

The greedy algorithm helps to improve the model performance after adding f_t. After that, the output of the model in each iteration by minimizing the objective function is given as (Alabdullah et al., 2022; Liu et al., 2023):

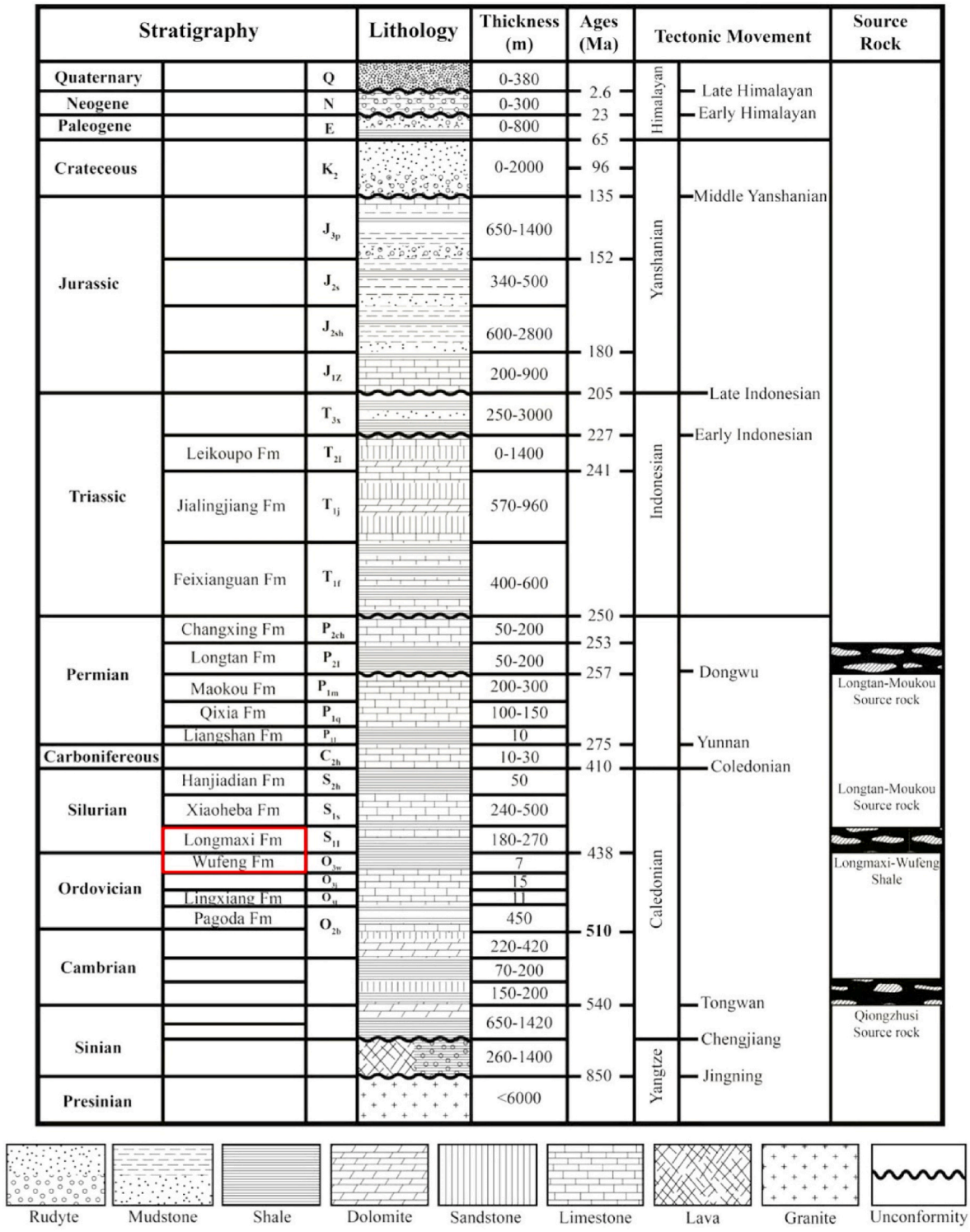


Fig. 2. The Stratigraphic column of the Southern Eastern Sichuan Basin shows the stratigraphic occurrence of Upper Ordovician Wufeng to Lower Silurian Longmaxi shale gas reservoir (Mgimba et al., 2023a).

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(X_i) \tag{6}$$

4.2. Light gradient boosting machine (LightGBM)

LightGBM is a machine learning algorithm belonging to the Gradient Boosting Decision Trees (GBDT) family. It is based on a decision-tree approach introduced by Microsoft Research (Sun et al., 2021). The proposal of LightGBM emerged as a solution to address the limitations of typical GBDT methods, particularly in handling large-scale datasets (Hajihosseinlou et al., 2023). LightGBM successfully approaches regression, classification, and other machine-learning problems. Achieving accuracy in forecasting entails a lower memory need. Its goal is to improve computing efficiency and solve difficulties in large-scale prediction (Liang et al., 2020b). The LightGBM integrates two cutting-edge approaches: gradient-based one-side sampling (GOSS), which handles a huge data set, and exclusive features bundling (EFB),

Table 3
Statistical analysis of the training data.

Features	K (MPa)	DTC ($\mu s/m$)	DTS ($\mu s/m$)	v	R (Ωm)	G (MPa)	E (MPa)	BI (%)
Mean	35.7061	60.6947	100.8989	0.2136	1.6655	24.7158	59.9443	69.9761
Standard deviation	8.4732	5.0163	7.3590	0.0384	0.0788	3.4788	8.3043	9.4509
Minimum	15.352	48.017	86.945	0	1.378	14.633	35.794	42.504
Maximum	71.526	76.307	124.677	0.33	1.986	32.645	78.649	91.147
Range	56.174	28.29	63.601	0.33	0.608	18.012	42.855	48.643

Table 4
Statistical analysis of the testing data.

Features	K (MPa)	DTC ($\mu s/m$)	DTS ($\mu s/m$)	v	R (Ωm)	G (MPa)	E (MPa)	BI (%)
Mean	46.4953	58.6262	106.9269	0.2833	1.8327	22.5649	57.9988	64.1728
Standard deviation	15.1056	7.3670	10.6267	0.0355	0.1065	4.5192	12.1689	25.6444
Minimum	13.679	45.338	88.013	0	1.378	7.631	19.838	24.539
Maximum	83.587	81.884	151.614	0.367	2.184	33.04	86.156	111.044
Range	69.907	36.546	63.601	0.367	0.806	25.409	66.318	86.505

which manages many data features without causing overfitting issues. Both of these techniques are designed to manage data more effectively. LightGBM uses the histogram technique and the tree leaf-wise growth strategy, contributing to its enhancements in computing efficiency and predictive accuracy (Guo et al., 2023). It combines weak learners to form strong models' that can be used for prediction. Its objective is to minimize the loss function, commonly formulated as an overall sum of the losses incurred by each case within the dataset. The loss function can be expressed as (Omotehinwa et al., 2023);

$$L(\theta) = \sum_{i=1}^{n} l(y_i, f(x_i, \theta)) + \Omega(\theta) \tag{7}$$

Where $L(\theta)$ is the loss function required to be minimized with respect to the parameter θ. The first term $\sum_{i=1}^{n} l(y_i, f(x_i, \theta))$ is empirical risk, which measures the differences between true output (y_i) and predicted output $f(x_i, \theta)$ for individual training samples. The second term $\Omega(\theta)$ is the regularization term, which helps to prevent overfitting. LightGBM offers various choices for specifying the type of regularization to be employed. The parameter denoted as 'lambda' governs the strength of L2 regularization, whereas the parameter denoted as 'alpha' governs the strength of L1 regularization (Omotehinwa et al., 2023). For the benefit of the readers, the difference between LightGBM and XGBoost is shown in Table 5. The similarities between the methods include the following: both models' can be applied for regression and classification problems; both models' let you choose your own training goals and assessment criteria; both methods provide numerous hyperparameters tuning to reduce overfitting and increase model generalizations; both models' combine weak learners to form strong learners models' which can be used for prediction purpose etc.

4.3. K-nearest neighbor (KNN)

K-Nearest Neighbors (KNN) is a simple and commonly used machine learning classification and regression technique. It is classified as instance-based learning or lazy learning because it is flexible to new data as it incorporates the new instances data into the decision-making process without retraining the model again; there is no explicit training phase because the algorithm stores the training instances for later use, direct use of instances, i.e., KNN directly uses the training instances as its "knowledge." It doesn't try to summarize the data into a model representation. It treats each training instance as a separate piece of information that can be consulted when needed (Ghunimat et al., 2023; Wang et al., 2023a). The basic principle of KNN is to classify or forecast the output of a new data point in the feature space based on the majority class (for classification) or average (for regression) of its k nearest neighbors (Jodas et al., 2023; Uddin et al., 2022). The fundamental principle of KNN is that identical and completely distinct samples are separated by shorter distances in the high-dimensional mapping space. The distance is calculated using the Euclidian formula, as shown in Eq. (8) (Kurniadi et al., 2018). The average values of the K nearest neighbors' targets are determined based on the characteristics of true values when the input data enters the KNN model.

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \tag{8}$$

Where d_i represents the distance of data variables, x_{2i} is sample data, x_{1i} stands for testing data. The steps for KNN algorithms executions are: 1) Preparing data sample in the form of an array. 2)Preparing testing data in the form of an array; 3) Computing Euclidian distance between testing data; 4) Separating the distance results based on the lowest values and a predetermined number of neighbors; 5) Obtain predicted outputs based on the calculation of the highest number using Eq. (9) (Fan et al., 2019); 6) Computing the accuracy based on the prediction results (Kurniadi et al., 2018).

$$s_i = \frac{1}{k} \sum_{j=1}^{k} S_{y_j} \tag{9}$$

Where s_i stands for the ith predicted value, which is the average value of $S_{y_j} (j = 1, 2, \ldots, k)$; S_{y_j} stands for the predicted value of the *jth* closest known data point (y_j).

4.4. Particle swarm optimization-random forest (PSO-RF)

Particle swarm optimization-random forest (PSO-RF) combines particle swarm optimization (PSO) with the random forest (RF) algorithm. The idea behind this hybrid is to leverage the strengths of both techniques to achieve better performance in certain scenarios (Chatrsimab et al., 2020; Shi and Zhang, 2023). However, whether PSO-RF consistently performs better than RF depends on the specific problem, dataset, and tuning parameters involved (Grichi et al., 2018; Liang et al., 2020a; Wang et al., 2023b).

4.4.1. Particle swarm optimization

Particle Swarm Optimization (PSO) is an optimization population-based technique that operates on a swarm of particles, drawing inspiration from the collective behaviour observed in birds and fish (Wang et al., 2022). In PSO, a collective of particles representing potential solutions undergoes iterative alterations to their placements, using both individual experiences and the collective experiences of the entire group. The objective is to determine the optimal solution by

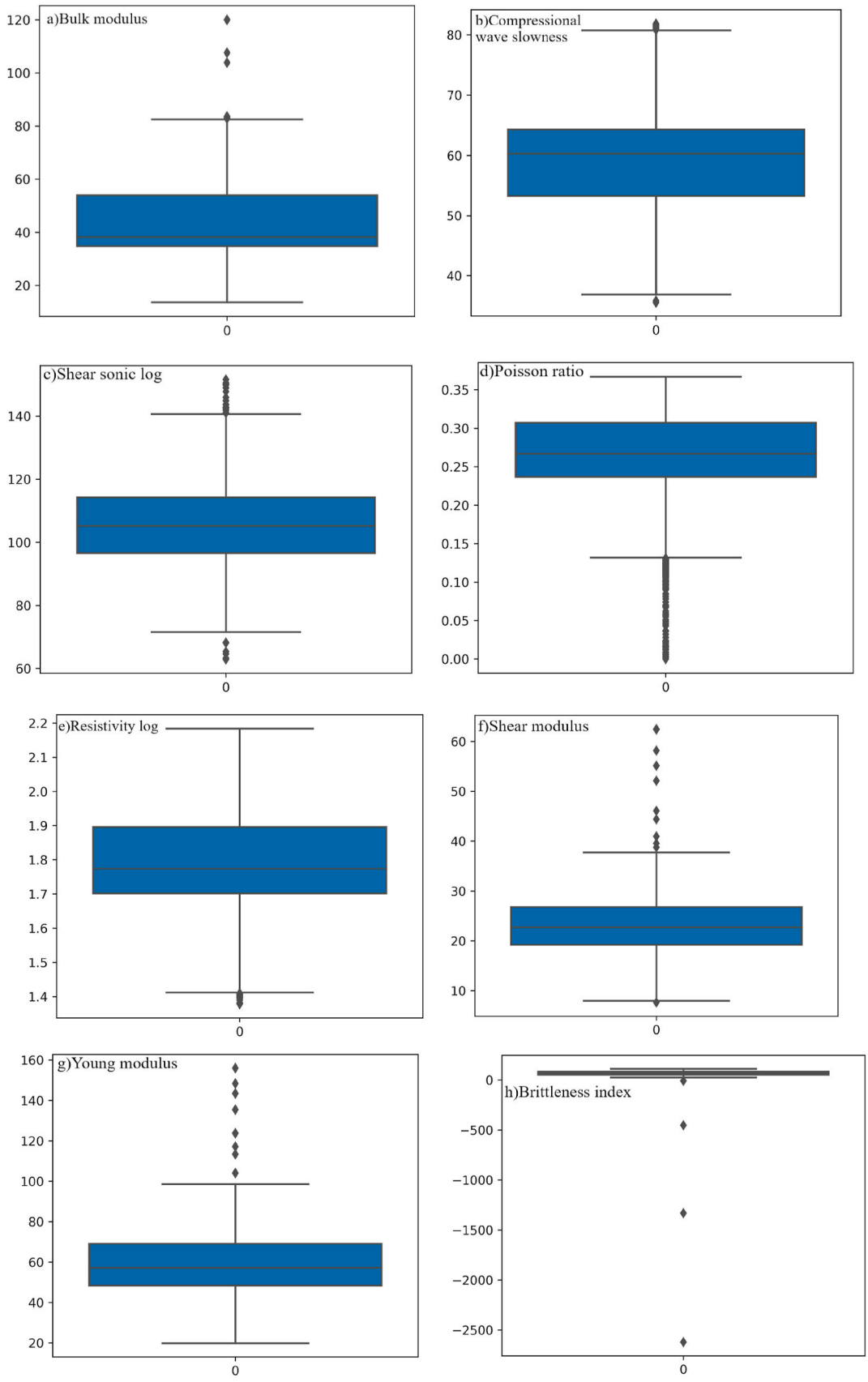


Fig. 3. Combined datasets with outliers.

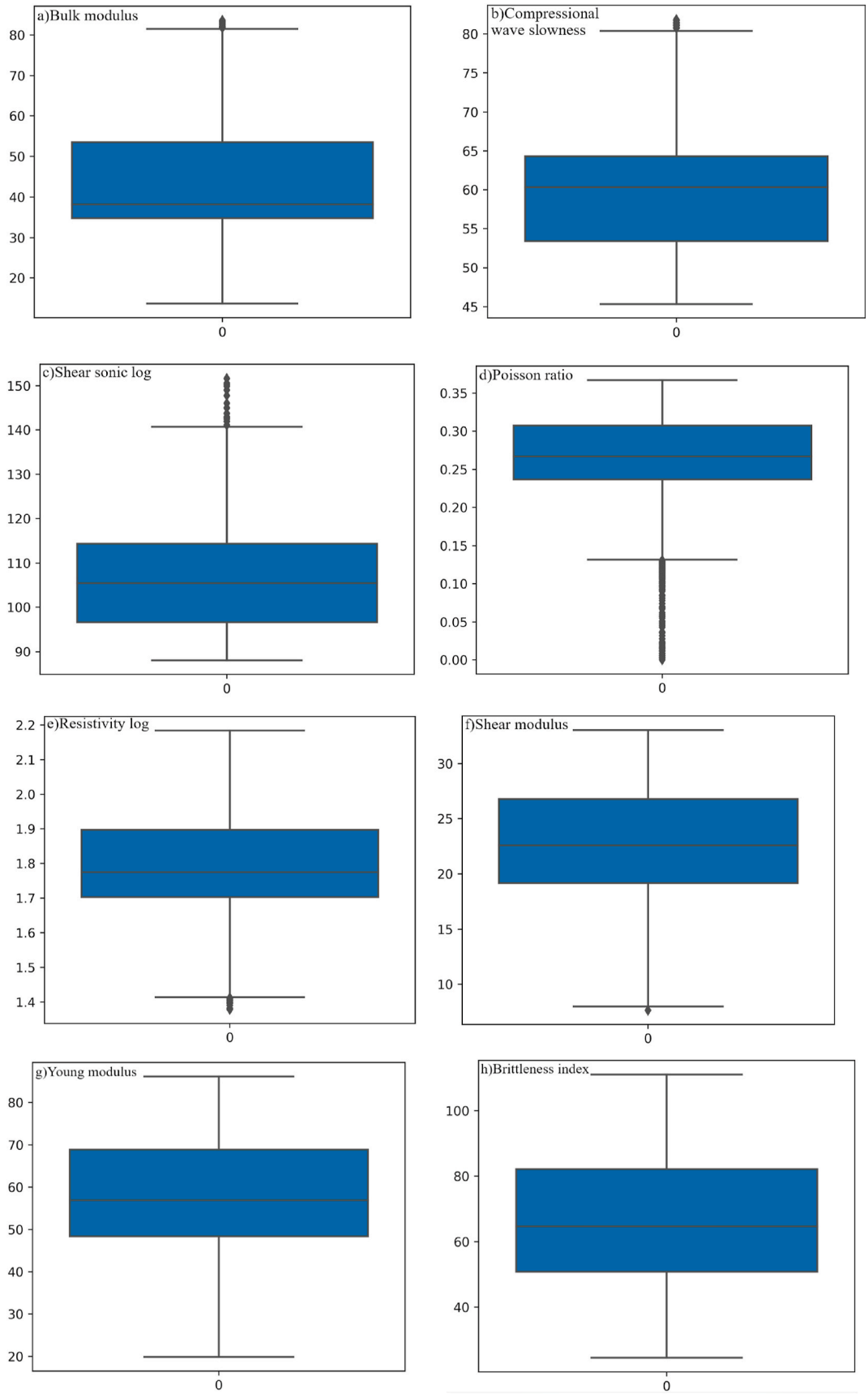


Fig. 4. Combined datasets without outliers.

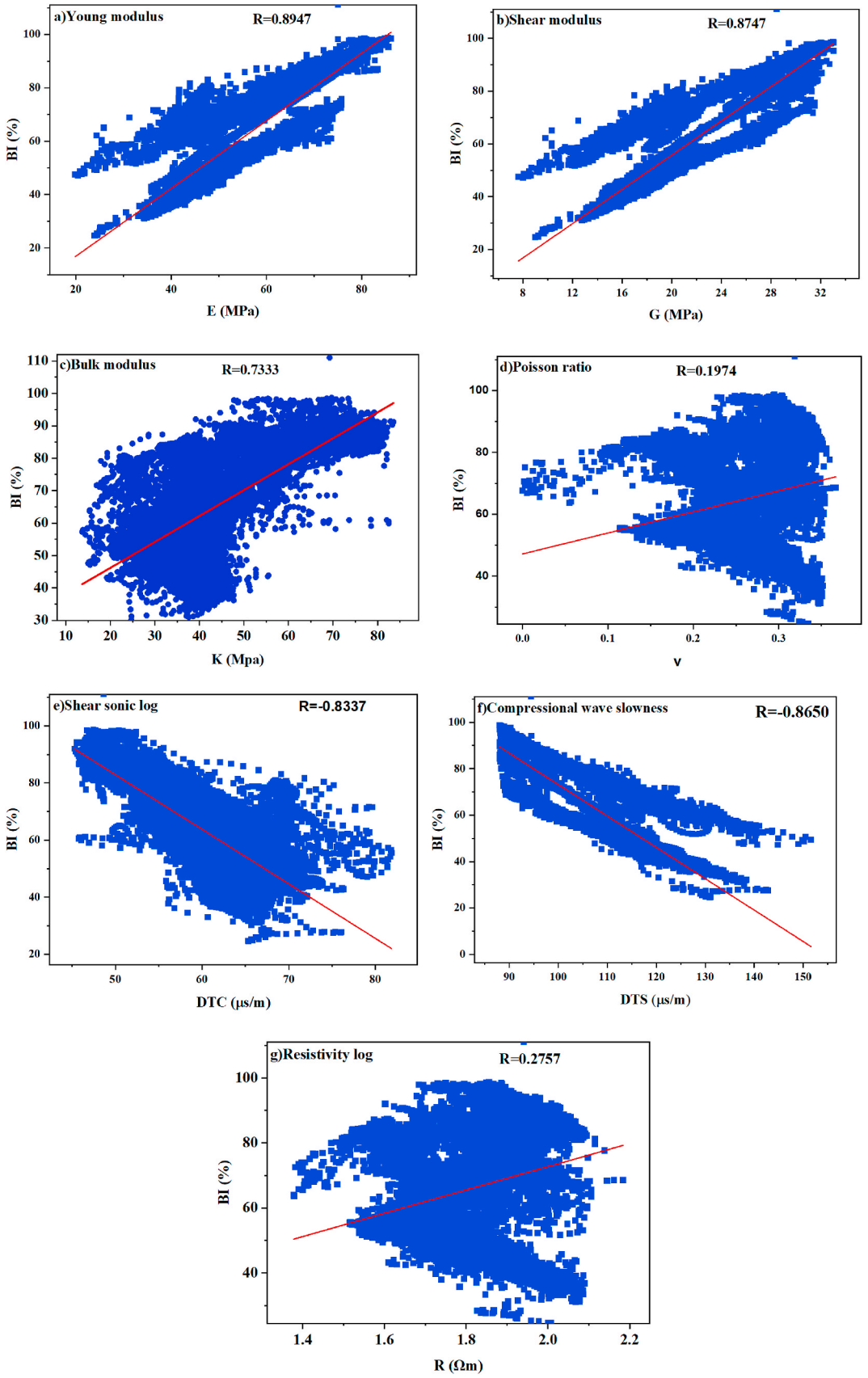


Fig. 5. Pearson correlation between Brittleness index with inputs.

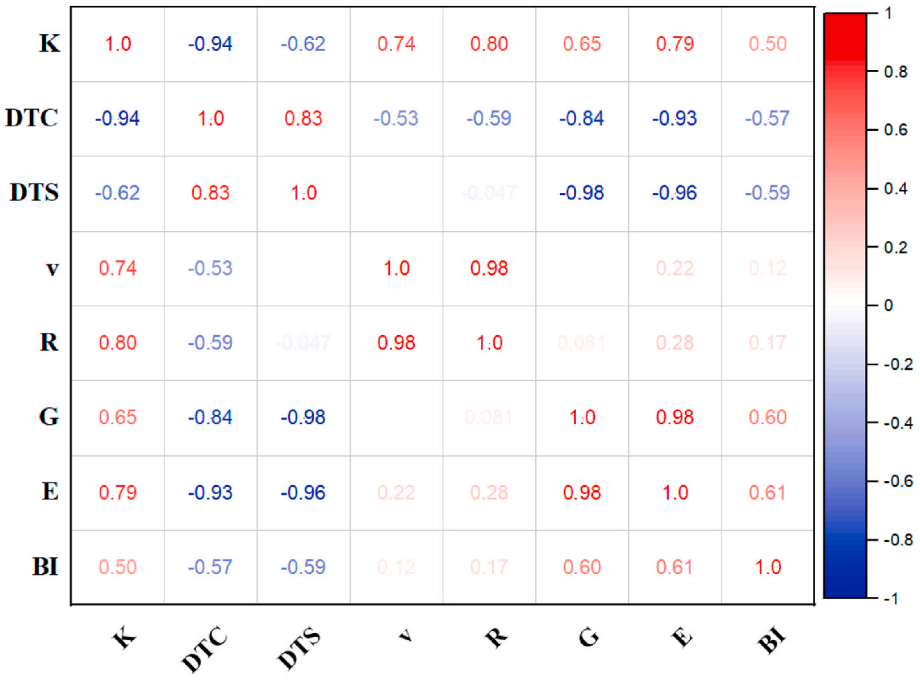


Fig. 6. Correlation coefficient matrix between the datasets.

Table 5
Difference between LightGBM and XGBoost.

Parameter	LightGBM	XGBoost
Growth strategy	LightGBM employs a leaf-wise tree growth technique, where the algorithm selects the split that yields the greatest improvement to the objective function. A deeper and more complex tree structure may result from this method.	XGBoost utilizes a level-wise tree growth technique in which each level of the tree is extended at the same time. Because of its shallow trees, it may be less prone to overfitting.
Missing value handling	LightGBM can handle missing values effectively throughout training and prediction without explicit imputation.	XGBoost requires missing value imputation before training.
Parallelization and speed	LightGBM is recognized for having a faster training speed because it uses a histogram-based method that lets more parallelization happen.	XGBoost's parallelization has improved, although LightGBM is usually deemed faster in training speed.
Memory usage	LightGBM's histogram-based and feature bundling techniques make it particularly memory-efficient when working with enormous data sets.	XGBoost could use more memory than LightGBM, particularly when working with large datasets.
Handling categorical features	LightGBM supports categorical characteristics by default. It uses integer indices to represent categories; therefore, it can handle categorical data without one-hot encoding.	XGBoost supports categorical features but needs one-hot encoding. LightGBM's category feature processing is faster and memory-efficient.

manipulating particles towards the optimal point within the search space. Suppose a population space with N dimension with each particle having an initial velocity and a position vector. The particle position and velocity are changed many times until the optimal solution is obtained (Deng and Jia, 2022).

Based on the particle principle behaviour, the PSO model is developed. The positions and velocity of the particles are expressed as $X_i = (x_{i1}, x_{i2}, ..., x_{iN})^T$ and $V_i = (V_{i1}, V_{I2}, ..., V_{iN})^T$, respectively. At the same time, the individual population extremum is expressed as $P_i = (P_{i1}, P_{i2}, ..., P_{iN})^T$ and the global population extremum is expressed as $P_g = (P_{g1}, P_{g2}, ..., P_{gN})^T$. Then, the iteration formula of particles is expressed as (Grichi et al., 2018; Wang et al., 2022, 2023b):

$$V_{in}^{k+1} = \omega V_{in}^{k} + c_1 r_1 \left(P_{in}^{k} - X_{in}^{k}\right) + c_2 r_2 \left(P_{gn}^{k} - X_{in}^{k}\right) \tag{10}$$

$$X_{in}^{k+1} = X_{in}^{k} + V_{in}^{k+1} \tag{11}$$

Where ω represents the inertia weight, k represents the maximum iterations, V_{in} is velocity, c_1 and c_2 stands for constants, r_1 and r_2 are distribution within $[0, 1]$. Position and speed are always limited within a certain range, enabling the particles to search quickly.

4.4.2. Random forest

Random forest is a popular ensemble ML algorithm that belongs to the family of decision tree-based and is designed to improve the accuracy and robustness of individual decision trees. It can be used to solve regression and classification problems. It integrates many decision trees during processing, and then the final results are derived by merging the outcomes of individual decision trees (He et al., 2023; Rigatti, 2017; Saha et al., 2022). The bootstrap sampling strategy is used to acquire the training samples linked to every base learner in the RF algorithm. Inother words, a subset is picked randomly from all the characteristics, and the samples left behind after the subset is drawn are called out-of-bag samples (OOB). The training samples and feature vectors of the tree are randomly generated with replacement. Both of these elements show the randomness of the tree. Because of this, it is possible to circumvent the overfitting issue, and the unpredictability of the training extraction process contributes to a greater degree of difference across the various decision trees (Cutler et al., 2012; Fan et al., 2022; He et al., 2022; Zhou et al., 2020). At this point, the forest has been constructed, and the results are computed by Eq. (12). Therefore, the final model formed has better accuracy (Grichi et al., 2018; Wang et al., 2022, 2023b).

$$\bar{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T(x, O_b) \quad (12)$$

Where $\bar{f}_{rf}^{B}$ is the average tree output, B represent the trees, and $T(x, O_b)$ represents the output of each tree.

The stages involved in the PSO-RF model development include.

Step 1. After removing outliers in well datasets (WY1) and (WY2), all the brittleness index (BI) data with associated inputs such as young modulus (E), bulk modulus (K), shear modulus (G), compressional wave slowness (DTC), shear sonic log (DTS), resistivity log (R), and poison ratio (v) were standardized from zero to one.WY1 datasets were used to train the PSO-RF model, whereas WY2 datasets were used to test the model's accuracy.

Step 2. Initialize PSO parameters: This involves defining the hyperparameters for RF, such as n_estimators, max_depth, min_samples_leaf, max_features, etc., then followed by setting PSO parameters such as number of particles, maximum iterations, inertia weight, etc.

Step 3. Initialize particle swarm: At this stage, a swarm of particles is generated, each representing sets of RF hyperparameters. After that, particle positions and velocities within the hyperparameters space are initialized.

Step 4. Evaluate initial particle positions by evaluating each particle performance (RF configuration) using cross-validation. Then, after each particle's best position and performance are updated.

Step 5. Main PSO loop: For each iteration, particle velocities must be updated, followed by evaluating particle positions, and then each particle and global bests are updated. Updating particle velocities involves calculating new velocities of the particle based on the historical and the swarm's global best and applying acceleration coefficients and inertia weight to balance exploitation and exploration. After that, the calculated velocities are used to update particle positions, and it is important to ensure that each parameter is defined within hyperparameter space. Then, the particle positions are evaluated using cross-validation. Updating the personal and global best scenario involves observing if the performance of the particle is better than its personal best. The personal best needs to be updated. If the performance of the particle is better than the swarm global best, then the global best needs to be updated.

Step 6. PSO-RF finalization: After the predefined number of iterations or convergence is met, the RF configuration associated with the global best particle is selected, followed by a training RF model utilizing selected configurations on the training data.

Step 7. PSO-RF model evaluation: This involves assessing the trained model using testing/validation data by observing models' performance indicators, i.e., R^2, RMSE, and MAE used for this paper. If the results are not good, steps 2–6 are repeated until the best results are obtained.

Step 8. Output the optimized RF model: After getting the best output, the RF model with optimized hyperparameters is obtained. The developed model can be applied for new data having inputs but missing the output to test its validity. This paper used the developed model to predict the BI index for two pilot test wells.

Step 9. End; the predicted outputs are obtained using the PSO-RF model. The flowchart for PSO-RF is shown in Fig. 7.

4.5. Hyperparameter tuning in models development

Hyperparameter tuning is an important step in ML to ensure that the models' predict the outputs of unseen data through algorithm optimization. Tuning these hyperparameters correctly affects the model output

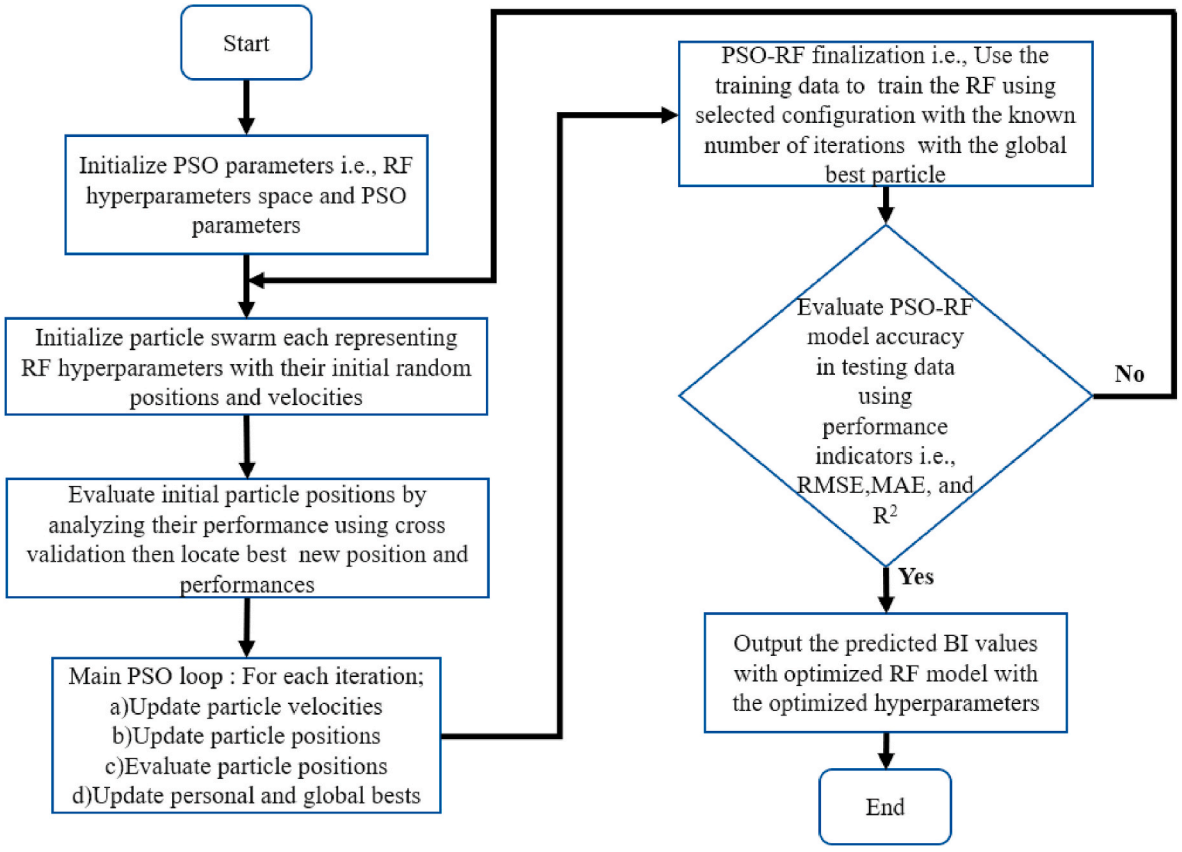


Fig. 7. Flowchart for generalized Random Forest based on particle swarm optimization (PSO-RF).

(Bischl et al., 2023; Feurer and Hutter, 2019; Yang and Shami, 2020). In this paper, the random search was used for hyperparameter optimization for PSO-RF, XGBoost, LightGBM, and KNN to enhance the accuracy of the models'. The random search involves randomly sampling hyperparameter combinations from predefined ranges. This method can be more efficient than grid search because it explores different combinations without exhaustively searching all possibilities. It evaluates performance for all combinations of hyperparameters and their values and finds the best value. The hyperparameter optimization via random search for this study is shown in Table 6.

5. Results and discussions

5.1. Models' performance indicators

Performance indicators are measurements used to evaluate the efficiency and effectiveness of ML models'. These indicators help data scientists and ML practitioners to understand the strengths and weaknesses of their models' by providing insights into how a model works and assisting them in understanding the strengths and weaknesses of their models'. In this paper, Python 3.11.5 version software was used for models' development. Three model performance indicators were used to assess the model efficiency and effectiveness in predicting BI, which are coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), as shown in Eqs. (13)–(15), respectively (Mkono et al., 2023; Mulashani et al., 2022). The coefficient of determination quantifies the strength and direction of a linear relationship between two variables. It determines the degree of agreement or similarity between the predicted and actual values. Mean absolute error quantifies the average absolute difference between predicted and actual values. One of the advantages of MAE is that it treats all errors equally, regardless of their magnitude, making it a good choice when outliers are present in the data. However, it doesn't indicate the direction of errors or whether the model tends to overestimate or underestimate the actual values. A lower RMSE indicates better predictive performance, which signifies that the model's predictions are closer to the actual values. A higher RMSE indicates poorer predictive performance, which signifies that the model's predictions are farther from the actual values. It has been reported that the model performance is excellent when it gives R^2 value closer to 1 for training and testing. Also, If the RMSE and MAE are close to zero or significantly smaller than the range of the target variable, it indicates a good model fit (Dabiri et al., 2022; Wood, 2021).

$$R^2 = \frac{\left(\sum_{i=1}^{N}(y_{act} - \overline{y_{act}})(Y_{prd} - \overline{Y_{prd}})\right)^2}{\left(\sqrt{\sum_{i=1}^{N}(y_{act} - \overline{y_{act}})^2}\right)\left(\sqrt{\sum_{i=1}^{N}(Y_{prd} - \overline{Y_{prd}})^2}\right)^2} \tag{13}$$

$$RMSE = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(y_{act} - Y_{prd})^2\right)} \tag{14}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_{act} - Y_{prd}| \tag{15}$$

Where y_{act} is actual value, $\overline{y_{act}}$ is average actual value, Y_{prd} is predicted value, $\overline{Y_{prd}}$ is the average predicted value, and N represents the quantity of data.

5.2. Models' statistical analysis

After training and testing the four models', the R^2, RMSE, and MAE were obtained. For training, R^2 were 0.9934,0.9796, 0.9519, and 0.9416 for PSO-RF, XGBoost, LightGBM, and KNN, respectively, as shown in Fig. 8. For testing, R^2 were 0.9533,0.9272, 0.9264, and 0.9262 for PSO-RF, XGBoost, LightGBM, and KNN, respectively, as shown in Fig. 8. A high value of R^2 during the training phase implies that the model effectively captures the underlying patterns present in the training dataset. Similarly, a high R^2 value obtained during the testing phase indicates that the model performs well when applied to unseen data. This indicates that the models' did not overfit training data. If there is a significant difference in the value of R^2 between training and testing is

Table 6
Optimized hyperparameters for the developed models'.

ML models'	Hyperparameters	Descriptions	Search space	Optimal values
KNN	n_neighbors	Number of neighbors	[1,25]	7
LightGBM	max_depth	Maximum depth of a tree	[2,20]	15
	n_estimators	Number of trees	[500,2000]	1800
	learning_rate	Shrinkage factor for each tree	[0.02,0.3]	0.1
	num_leaves	Number of leaves for each tree	[1100]	10
	min_data_in_leaf	Minimum number of data in leaf	[1,30]	18
	bagging_fraction	Number of events to be utilized for training a tree	[0.1,1.5]	0.8
	lambda_l1	Reduce the issue of overfitting	[0.1–2]	1
XGBoost	n_estimators	Number of trees	[500,2000]	1200
	max_depth	Maximum depth of a tree	[1,20]	12
	subsample	Number of subsamples for constructing trees	[0.1,1]	0.6
	colsample_by tree	Number of features or predictors used to train a tree	[0.1,1]	1
	learning_rate	Shrinkage factor for each tree	[0.01,1.5]	0.3
RF	n_estimators	Number of trees	[1000,2000]	1100
	max_depth	Maximum depth of a tree	[1,25]	22
	min_samples_leaf	Minimum number of samples for leaf nodes	[1,50]	38
	min_samples_split	Minimum number of samples for nodes split	[1100]	60

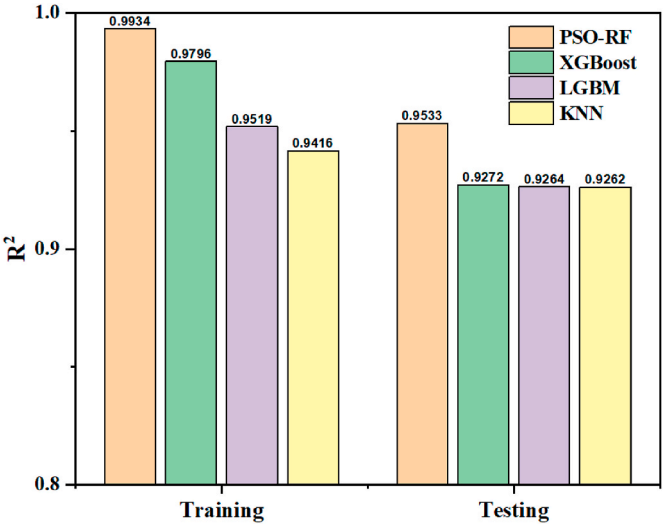


Fig. 8. Correlation coefficient comparisons for different models'.

when overfitting occurs, it captures noise and specific patterns that don't generalize well to testing data. It is crucial to consider both the training and testing R^2 values since they offer various perspectives on the effectiveness and generalization of the model. The model must perform well on training and testing data sets to achieve high generalization. Besides that, the RMSE values after training and testing all the models are shown in Fig. 9, in which PSO-RF had a minimum error compared to other models. In addition, the MAE values of PSO-RF were lowest compared to other models during training and testing, as shown in Fig. 10. The summary of the three performing indicators is shown in Table 7. After assessing statistical performance indicators, it has been revealed that PSO-RF outperformed XGBoost, LightGBM, and KNN models' during training and testing with R^2 of 0.9374, RMSE = 4.6327, MAE = 2.0974, and R^2 of 0.9329, RMSE = 15.5308, MAE = 5.3896, respectively. The order of performances for all models' in predicting BI was PSO-RF > XGBoost > LightGBM > KNN. The ability of PSO-RF to capture nonlinear relationships between many variables that are present in the dataset may be one of the reasons why it has a lower error rate and a better accuracy rate than other models. It may be difficult for other models', such as XGBoost, LightGBM, and KNN, to accurately predict the complex nonlinear interactions in the data, which might result in greater error rates. On the other hand, due to PSO-RF adaptability in terms of both the discovery and integration of such nonlinear correlations, it was able to make more accurate predictions for BI. In addition, PSO enables the optimization of both the feature selection and the hyperparameters of the RF model. PSO can assist in identifying a subset of pertinent features that exhibit a greater influence on the models' predictive capacity. This phenomenon enhances the generalization process and mitigates the issue of overfitting, particularly in scenarios characterised by a multitude of extraneous or duplicative variables.

5.3. Models' comparisons

As introduced in the methodology section, four models' were used in this paper. Hence, four ML models' were compared to assess their robustness and predictive accuracy in BI. This section analyzed cross plots to evaluate the models' performances. Cross plots, also known as scatter plots, are used to compare actual and predicted values in ML and statistics. When the points in a cross plot are close to the line y = x, the predicted values almost match the actual values, and the coefficient of determination is always high (close to one), indicating high accuracy of the used model. On the other hand, when the points in the cross plots are far away from the line y = x, the predicted values are not close to actual

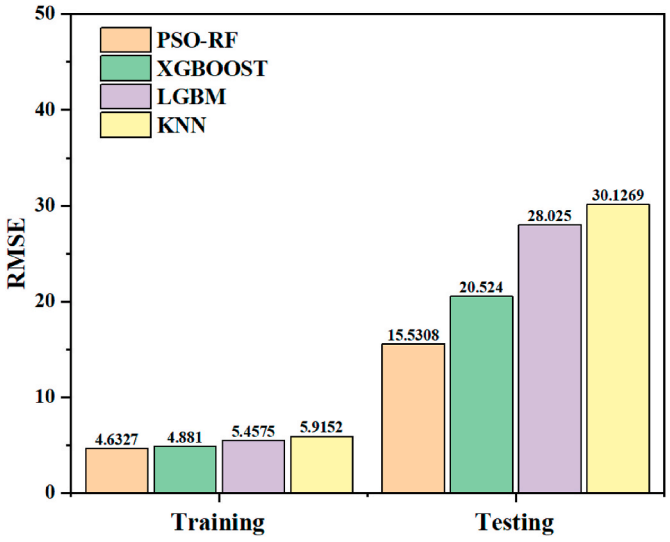


Fig. 9. RMSE comparisons for different models'.

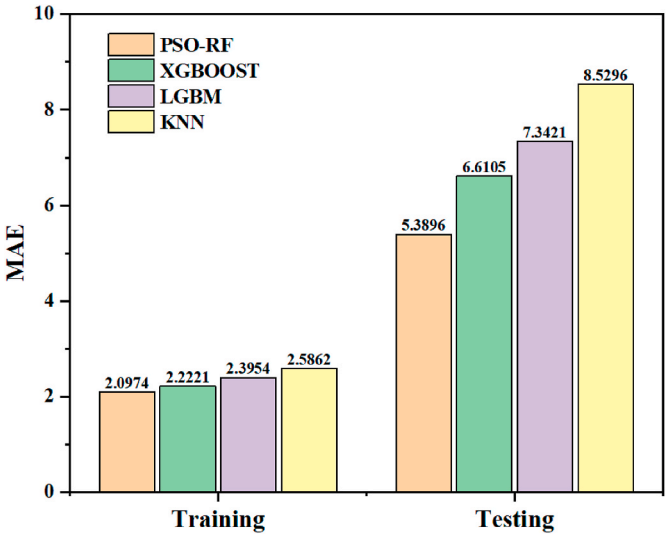


Fig. 10. MAE comparisons for different models'.

Table 7
Statistical results of the models'.

Model	R^2		RMSE		MAE	
	Training	Testing	Training	Testing	Training	Testing
PSO-RF	0.9934	0.9533	4.6327	15.5308	2.0974	5.3896
XGBoost	0.9796	0.9272	4.881	20.524	2.2221	6.6105
LightGBM	0.9519	0.9264	5.4575	28.025	2.3954	7.3421
KNN	0.9416	0.9262	5.9152	30.1269	2.5862	8.5296

values, indicating a low level of model accuracy with high errors. For the ideal scenario, the predicted values perfectly match the actual values with all the data points exactly falling on the line y = 1 with $R^2 = 1$. The cross plots for PSO-RF, XGBoost, LightGBM, and KNN during training are shown in Fig. 11 (a), 11 (b), 11(c), and 11 (d), respectively. For testing, the cross plots for PSO-RF, XGBoost, LightGBM, and KNN are shown in Fig. 12 (a), 12 (b), 12 (c), and 12 (d), respectively. Figs. 11 and 12 revealed that the BI values predicted by PSO-RF are better than the BI values predicted by XGBoost, LightGBM, and KNN during training and testing. The reason (s) behind this is that PSO-RF has PSO, which helped optimize RF hyperparameters for each particle in the swarm, increasing ensemble diversity and resulting in a more diverse ensemble of trees that improves model performances. Also, PSO explore and identify the optimal configurations of hyperparameters, such as the number of trees, tree depth, and minimum samples per leaf, which enhance model performance instead of employing default hyperparameters or manual adjusting settings. In addition, PSO helped in escaping local minima effectively as it avoids the model stuck in suboptimal configurations.

In addition, Taylor's diagram was employed to compare the predicted and actual BI. Taylor diagrams are type of graphical analysis tool that can be used to evaluate the accuracy of various models' or simulations in relation to a reference (actual) dataset. Taylor diagrams represent how well different models' replicate the observed data in terms of correlation and variability. Taylor's diagram consists of three major components: reference points representing the actual data; circular contours representing the standard deviation (SD), R^2, and model points. A model closer to the reference point better captures the observed variability and correlation. Furthermore, the angle between the model point, the reference point, and the origin represents the correlation between the model and the observations. Smaller angles indicate higher correlations. In addition, the distance along the circular contours from the reference point shows the SD of the model's diversity. Models' with similar SD to the reference data will be more accurate

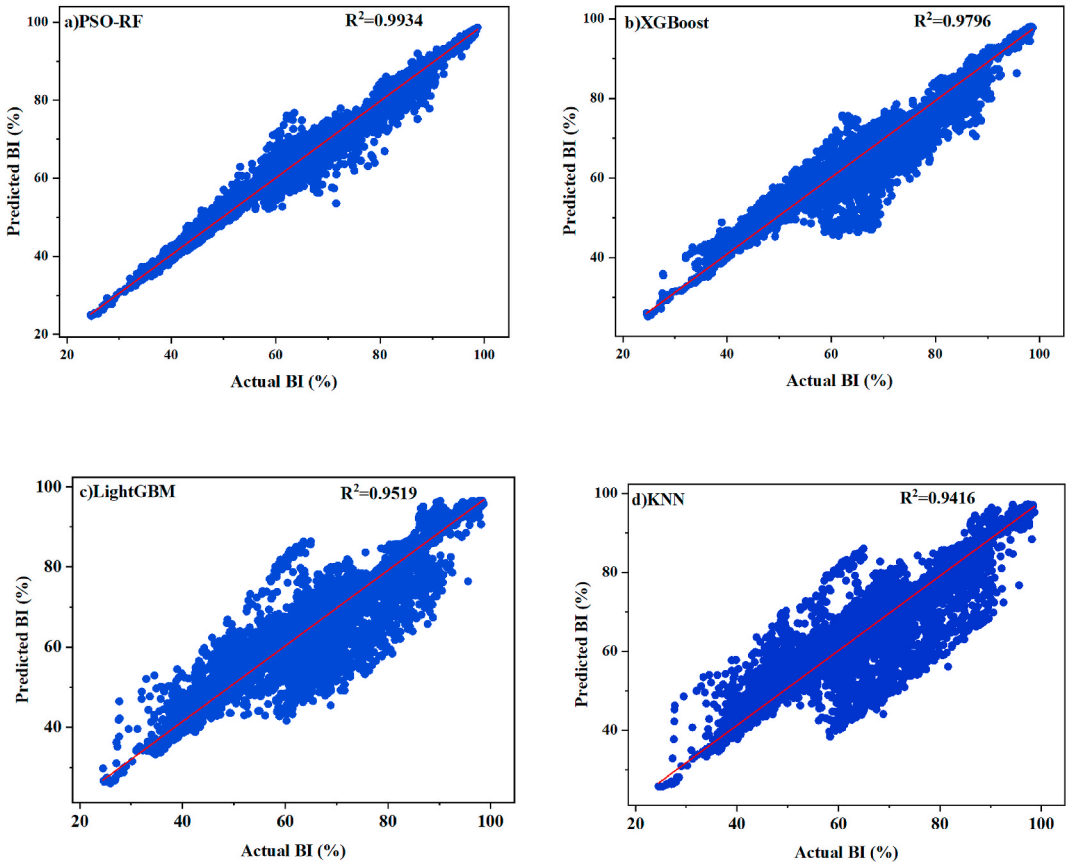


Fig. 11. Cross plots for the predicted versus actual BI in training WY1 datasets.

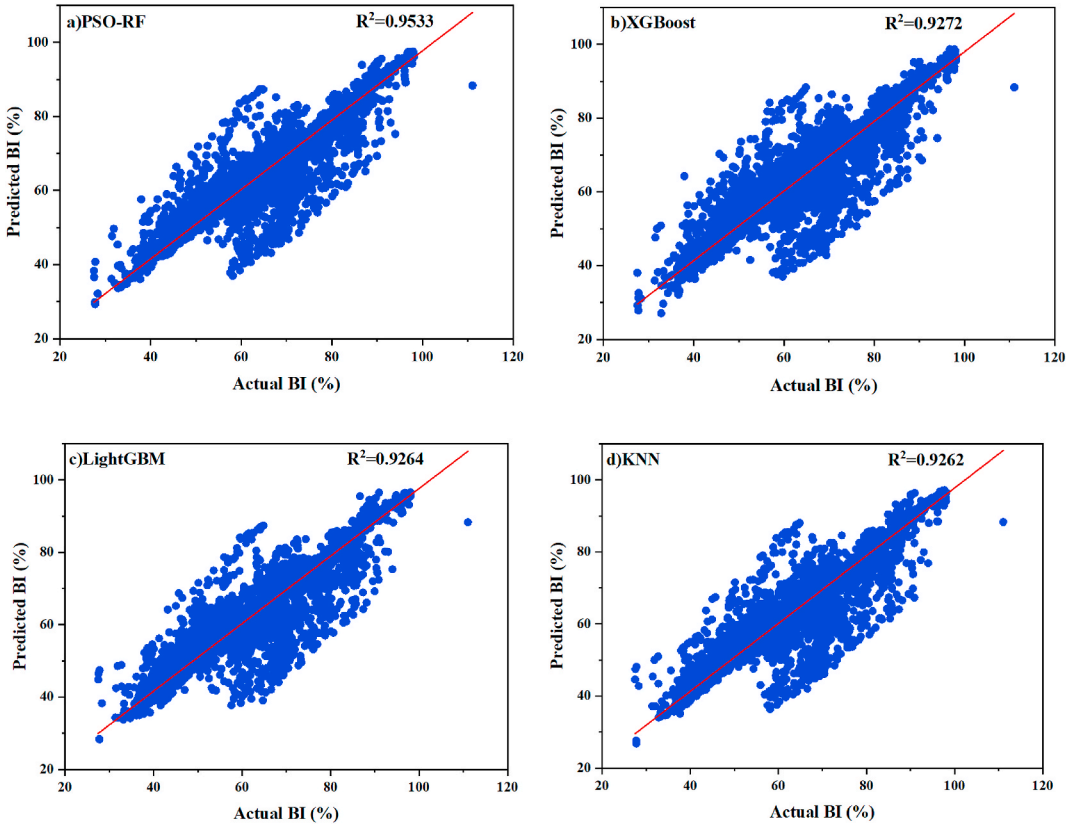


Fig. 12. Cross plots for the predicted versus actual BI in testing WY2 datasets.

(Dabiri et al., 2022; Gleckler et al., 2008; Kioumarsi et al., 2023; Xi et al., 2023). From Fig. 13, it has been shown that PSO-RF outperformed Adaboost, XGBoost, and KNN in the prediction accuracy of BI as its SD and R^2 values are closer to actual values. The models' performance assessment through Taylor's diagram agrees with other model assessment methods, i.e., PSO-RF > XGBoost > LightGBM > KNN.

5.4. SHAP analysis

SHAP (SHapley Additive exPlanations) is a popular method for interpreting machine learning models' by quantifying the contribution of each input feature to the model's output. It provides a way to understand the relationship between individual features and the predicted outcome (Alabdullah et al., 2022; Mangalathu et al., 2020; Zhang et al., 2023). SHAP analysis is not limited to any specific type of machine learning model; it can be used with various models', including linear regression, decision trees, random forests, and gradient boosting. The most common way to analyze the correlations between input features and the output using SHAP values is to create a SHAP summary plot or a feature importance plot. In this scenario, each feature is represented as a horizontal bar. The bar's position on the plot indicates the average SHAP value for that feature across the dataset, and the colour of the bar shows whether the feature value is high (red) or low (blue) (Ji et al., 2022; Yang et al., 2021). Analyzing the SHAP summary plot lets you gain insights into how each feature affects the model's predictions. Features with larger positive SHAP values push the predictions higher, while features with larger negative SHAP values push the predictions lower. Features closer to the baseline prediction have less impact on the model's output (Mangalathu et al., 2020; Nazar et al., 2023). From Fig. 14, it has been revealed that E, G, and K are the most important input parameters affecting the BI of shale gas formations, whilst v and R are the least input parameters affecting the BI of the shale gas formations. Besides that, Fig. 15 shows that the increase in E and G influences the increase in BI of the formations. Further, the decrease in DTS and DTC results in the increase of BI of the formation. In addition, the high value of R and v decreases BI.

5.5. Validation of the results

After the successful application of PSO-RF in predicting the BI of Weiyuan wells (WY1 and WY2) of the Sichuan Basin with high accuracy, the developed model was used to predict the BI of another two new pilot test wells (WY3 and WY4), which were not used to build the model. Well

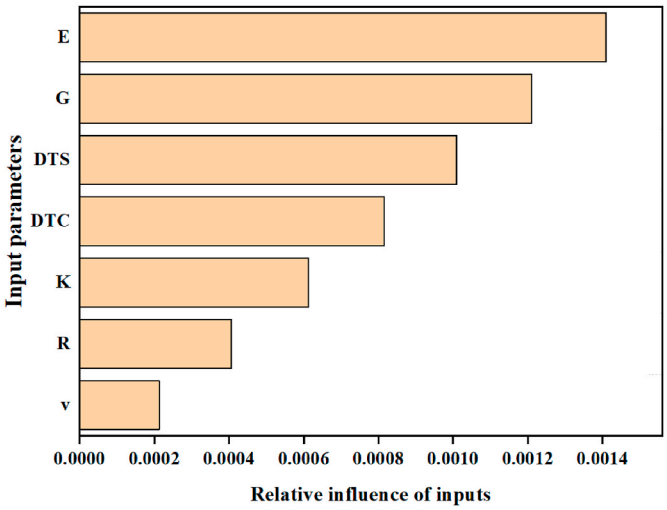


Fig. 14. Relative influence of inputs to BI using PSO-RF.

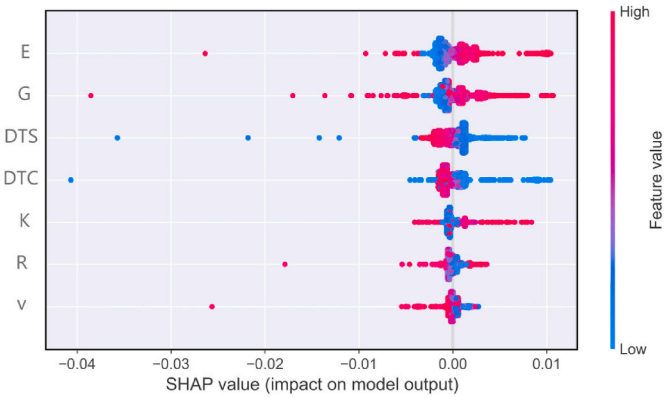


Fig. 15. SHAP analysis of the inputs to BI using PSO-RF.

WY3 had the same inputs as wells WY1 and WY2, such as E, K, G, v, DTC, DTS, and R. Further, wells WY3 had BI data, while WY4 had no BI data. For well WY3, we assumed that BI does not exist and its inputs are used to predict new BI. The correlation plot between the field and predicted BI of the well WY3 is shown in Fig. 16, which shows that the coefficient

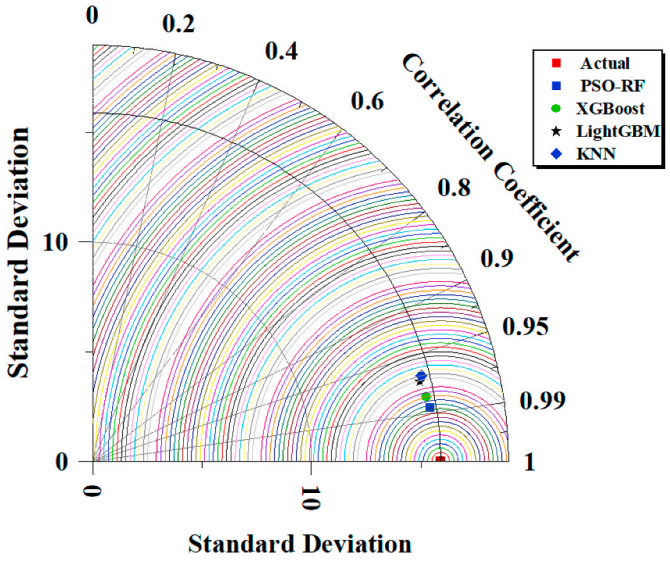


Fig. 13. Models' evaluation using Taylor's diagram.

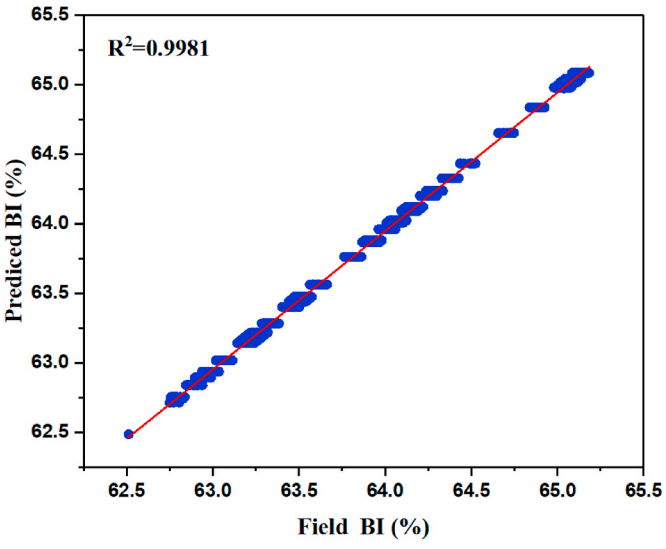


Fig. 16. Cross plots between predicted and field BI for well WY3.

of determination is 0.9981 while RMSE is 0.057699, and MAE is 0.01499. In general, the coefficient of determination between the field and predicted BI is close to one, which validates the model accuracy results. Further, RMSE and MAE are close to zero, validating the developed model results. For well WY4, which had no BI values, the developed model was used to predict the BI, which remained very stable throughout the entirety depth of the wells, as shown in Fig. 17. Because of the nearly consistent BI values over the whole depth of the well, the predicted BI using PSO-RF shows that the potential depth for hydraulic fracturing operation is marked with red boxes (Fig. 17) because low brittleness formation limits the vertical growth of fractures, thus prevent the fractures to penetrate the potential hydrocarbons zone for production purpose. According to Fig. 17, the formation for hydraulic fracturing operations is the Wufeng formation (First red box) and Linxiang (second red box) formations from the top. However, BI is not the only parameter to determine the possible layers for hydraulic fracturing in shale gas reservoirs. Many parameters such as fracability, fracture toughness, tensile strength, etc., can be grouped with BI to help make decisions on hydraulic fractures placement.

6. Conclusions

This study used Random Forest based on particle swarm optimization (PSO-RF) to develop a novel model for predicting the brittleness index (BI) of Upper Ordovician Wufeng to Lower Silurian Longmaxi shale gas formation from Young modulus (E), bulk modulus (K), shear modulus (G), compressional wave slowness (DTC), shear sonic log (DTS), resistivity log (R), and poisson ratio (v). The model predicted BI with high accuracy and the least errors compared to Extreme gradient boosting (XGBoost), Light gradient boosting machine (LightGBM), and K-nearest neighbor (KNN) models'. The following conclusions can be made based on the obtained results.

1) During the training and testing phases, PSO-RF outperformed XGBoost, LightGBM, and KNN in BI prediction by yielding high R^2 and the least errors. PSO-RF R^2 were 0.9934 and 0.9533 during training and testing, respectively. RMSE and MAE were 4.6327 and 2.0974, and 15.5308 and 5.3896, during training and testing, respectively. The performance order of the models' was PSO-RF > XGBoost > LightGBM > KNN. This confirms that optimized ensembles can predict BI better than individual ensemble methods.
2) From SHAP analysis, it has been found that the Young modulus (E) and Shear modulus (G) of the shale gas reservoirs greatly influence BI, with resistivity log (R) and poisson ratio (v) having the least influence. E and G increase results in BI increase, while the decrease of R and v increases BI of the shale gas formations.
3) After the proposed PSO-RF model was applied to predict BI of the other two wells for model validation results, it was found that the model predicted BI with great accuracy and helped to locate where hydraulic fractures can be placed to enhance shale gas production from the Weiyuan gas field.

This study only focused on BI as one of the important parameters in determining the pay zone for hydraulic fracturing operations, whilst there are many parameters besides BI to be considered. Hence, we recommend future research to consider other parameters in their study, such as fracability, fracture toughness, tensile strength, etc. In addition, the proposed model in this research may also be used to forecast other hydraulic fracturing parameters, such as fracture toughness, tensile strength, etc.

Credit authors statement

Mbula Ngoy Nadege: Conceptualization, Writing – original draft, Writing – review & editing. Shu Jiang: Formal analysis, Validation. Grant Charles Mwakipunda: Validation, Formal analysis, and Visualization. Allou Koffi Franck Kouassi: Resources, Writing – review & editing, and Formal analysis. Paulin Kavuba Harold: Supervision, Formal analysis, Project administration, Resources, and Visualization. Konan Yao Hugues Roland: Formal analysis, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that there is no conflict of interest.

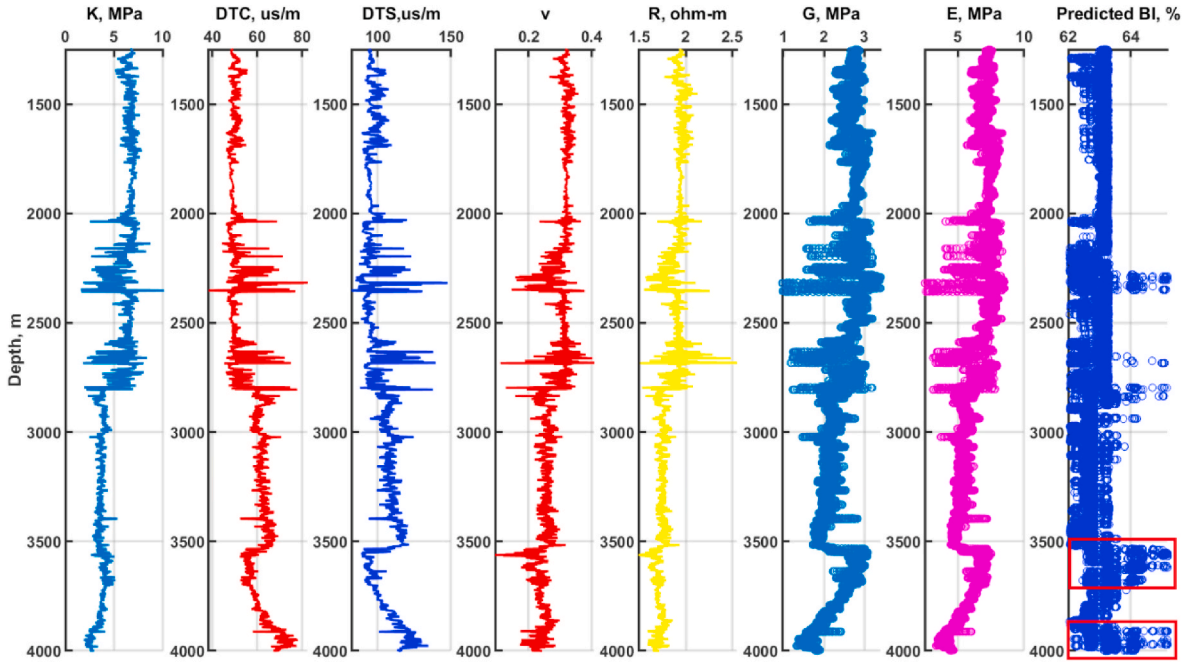


Fig. 17. Data inputs with predicted BI values using PSO-RF from well WY4.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank the support from National key research and development program (2022YFF0801201), National Natural Science Foundation of China (42130803), and Chinese Scholarship Council for their support.

References

Alabdullah, A.A., Iqbal, M., Zahid, M., Khan, K., Amin, M.N., Jalal, F.E., 2022. Prediction of rapid chloride penetration resistance of metakaolin based high strength concrete using light GBM and XGBoost models by incorporating SHAP analysis. Construct. Build. Mater. 345, 128296.
Andreev, G.E., 1995. Brittle Failure of Rock Materials. CRC press.
Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., et al., 2023. Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 13 (2), e1484.
Chatrsimab, Z., Alesheikh, A.A., Voosoghi, B., Behzadi, S., Modiri, M., 2020. Development of a land subsidence forecasting model using small baseline subset—differential synthetic aperture radar interferometry and particle swarm optimization—random forest (case study: tehran-karaj-shahriyar aquifer, Iran). In: Doklady Earth Sciences. Springer, pp. 718–725.
Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C.-H., Liu, R., 2019a. XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access 7, 13149–13158.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Chen, Z., Song, Y., Li, Z., Liu, S., Li, Y., et al., 2019b. The occurrence characteristics and removal mechanism of residual water in marine shales: a case study of Wufeng-Longmaxi shale in Changning-Weiyuan area, Sichuan basin. Fuel 253, 1056–1070.
Cornelio, J., Ershaghi, I., 2019. A Machine Learning Approach for Predicting Rock Brittleness from Conventional Well Logs. SPE Eastern Regional Meeting. SPE. D021S004R001.
Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random Forests. Ensemble Machine Learning: Methods and Applications, pp. 157–175.
Dabiri, H., Faramarzi, A., Dall'Asta, A., Tondi, E., Micozzi, F., 2022. A machine learning-based analysis for predicting fragility curve parameters of buildings. J. Build. Eng. 62, 105367.
Deng, J., Jia, G., 2022. An interpretable hybrid machine learning prediction of dielectric constant of alkali halide crystals. Chem. Phys. 555, 111457.
Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B., John, D., 2022. A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthc. Anal. 2, 100032.
Fan, G.-F., Guo, Y.-H., Zheng, J.-M., Hong, W.-C., 2019. Application of the weighted K-nearest neighbor algorithm for short-term load forecasting. Energies 12 (5), 916.
Fan, G.-F., Zhang, L.-Z., Yu, M., Hong, W.-C., Dong, S.-Q., 2022. Applications of random forest in multivariable response surface for short-term load forecasting. Int. J. Electr. Power Energy Syst. 139, 108073.
Feurer, M., Hutter, F., 2019. Hyperparameter Optimization. Automated Machine Learning: Methods, Systems, Challenges, pp. 3–33.
Gao, M., Li, T., Gao, Y., Zhang, Y., Yang, Q., He, Z., He, Q., 2023. A method to evaluation rock brittleness based on statistical damage constitutive parameters. Front. Earth Sci. 10, 1020834.
Ghunimat, D., Alzoubi, A.E., Alzboon, A., Hanandeh, S., 2023. Prediction of concrete compressive strength with GGBFS and fly ash using multilayer perceptron algorithm, random forest regression and k-nearest neighbor regression. Asian J. Civil Eng. 24 (1), 169–177.
Gleckler, P.J., Taylor, K.E., Doutriaux, C., 2008. Performance metrics for climate models. J. Geophys. Res. Atmos. 113 (D6).
Gong, F., Wang, Y., 2022. A new rock brittleness index based on the peak elastic strain energy consumption ratio. Rock Mech. Rock Eng. 55 (3), 1571–1582.
Grichi, Y., Dao, T.-M., Beauregard, Y., 2018. A new approach for optimal obsolescence forecasting based on the random forest (RF) technique and meta-heuristic particle swarm optimization (PSO). Proc. Int. Conf. Ind. Eng. Oper. Manag. 1680–1688.
Guan, Y., Yan, J., Shan, Y., Zhou, Y., Hang, Y., et al., 2023. Burden of the global energy price crisis on households. Nat. Energy 8 (3), 304–316.
Guo, J., Yun, S., Meng, Y., He, N., Ye, D., Zhao, Z., Jia, L., Yang, L., 2023. Prediction of heating and cooling loads based on light gradient boosting machine algorithms. Build. Environ. 236, 110252.
Hajihosseinlou, M., Maghsoudi, A., Ghezelbash, R., 2023. A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. Nat. Resour. Res. 1–22.
Hassan, A., Chan, S., Mahmoud, M., Aljawad, M.S., Humphrey, J., Abdulraheem, A., 2022. Artificial intelligence-based model of mineralogical brittleness index based on rock elemental compositions. Arabian J. Sci. Eng. 47 (9), 11745–11761.
He, B., Armaghani, D.J., Lai, S.H., 2023. Assessment of tunnel blasting-induced overbreak: a novel metaheuristic-based random forest approach. Tunn. Undergr. Space Technol. 133, 104979.
He, S., Wu, J., Wang, D., He, X., 2022. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere 290, 133388.
Huang, J., Caineng, Z., Jianzhong, L., Dazhong, D., Sheiiao, W., Shiqian, W., Cheng, K., 2012. Shale gas generation and potential of the lower Cambrian Qiongzhusi formation in the southern Sichuan Basin, China. Petrol. Explor. Dev. 39 (1), 75–81.
Hucka, V., Das, B., 1974. Brittleness determination of rocks by different methods. In: International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts. Elsevier, pp. 389–392.
Jamshidi, E.J., Yusup, Y., Kayode, J.S., Kamaruddin, M.A., 2022. Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: a case study on surface water temperature. Ecol. Inf. 69, 101672.
Jarvie, D.M., Hill, R.J., Ruble, T.E., 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 Bull. 91 (4), 475–499.
Ji, S., Wang, X., Lyu, T., Liu, X., Wang, Y., Heinen, E., Sun, Z., 2022. Understanding cycling distance according to the prediction of the XGBoost and the interpretation of SHAP: a non-linear and interaction effect analysis. J. Transport Geogr. 103, 103414.
Jodas, D.S., Passos, L.A., Adeel, A., Papa, J.P., 2023. PL-kNN: a Python-based implementation of a parameterless k-Nearest Neighbors classifier. Software Impacts 15, 100459.
Kavzoglu, T., Teke, A., 2022. Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). Arabian J. Sci. Eng. 47 (6), 7367–7385.
Khan, J.A., Padmanabhan, E., Haq, I.U., Franchek, M.A., 2023. Hydraulic fracturing with low and high viscous injection mediums to investigate net fracture pressure and fracture network in shale of different brittleness index. Geomech. Energy Environ. 33, 100416.
Kioumarsi, M., Dabiri, H., Kandiri, A., Farhangi, V., 2023. Compressive strength of concrete containing furnace blast slag; optimized machine learning-based models. Cleaner Eng. Technol. 13, 100604.
Kivi, I.R., Ameri, M., Molladavoodi, H., 2018. Shale brittleness evaluation based on energy balance analysis of stress-strain curves. J. Petrol. Sci. Eng. 167, 1–19.
Kivi, I.R., Zare-Reisabadi, M., Saemi, M., Zamani, Z., 2017. An intelligent approach to brittleness index estimation in gas shale reservoirs: a case study from a western Iranian basin. J. Nat. Gas Sci. Eng. 44, 177–190.
Kuang, Z., Qiu, S., Li, S., Du, S., Huang, Y., Chen, X., 2021. A new rock brittleness index based on the characteristics of complete stress–strain behaviors. Rock Mech. Rock Eng. 54, 1109–1128.
Kurniadi, D., Abdurachman, E., Warnars, H., Suparta, W., 2018. The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, 012039.
Lee, J., Lumley, D.E., 2023. Predicting shale mineralogical brittleness index from seismic and elastic property logs using interpretable deep learning. J. Petrol. Sci. Eng. 220, 111231.
Li, H., 2022. Research progress on evaluation methods and factors influencing shale brittleness: a review. Energy Rep. 8, 4344–4358.
Li, Y., Jia, D., Rui, Z., Peng, J., Fu, C., Zhang, J., 2017. Evaluation method of rock brittleness based on statistical constitutive relations for rock damage. J. Petrol. Sci. Eng. 153, 123–132.
Liang, J., Yan, C., Zhang, Y., Zhang, T., Zheng, X., Li, H., 2020a. Rapid discrimination of Salvia miltiorrhiza according to their geographical regions by laser induced breakdown spectroscopy (LIBS) and particle swarm optimization-kernel extreme learning machine (PSO-KELM). Chemometr. Intell. Lab. Syst. 197, 103930.
Liang, W., Luo, S., Zhao, G., Wu, H., 2020b. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. Mathematics 8 (5), 765.
Liu, B., Rostamian, A., Kheirollahi, M., Mirseyed, S.F., Mohammadian, E., Golsanami, N., Liu, K., Ostadhassan, M., 2023. NMR log response prediction from conventional petrophysical logs with XGBoost-PSO framework. Geoenergy Sci. Eng. 224, 211561.
Liu, W., Chen, Z., Hu, Y., 2022. XGBoost algorithm-based prediction of safety assessment for pipelines. Int. J. Pres. Ves. Pip. 197, 104655.
Majid, A., Mwakipunda, G.C., Guo, C., 2023. Solution gas/oil ratio prediction from pressure/volume/temperature data using machine learning algorithms. SPE J. 1–16.
Mangalathu, S., Hwang, S.-H., Jeon, J.-S., 2020. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Eng. Struct. 219, 110927.
Mangalathu, S., Karthikeyan, K., Feng, D.-C., Jeon, J.-S., 2022. Machine-learning interpretability techniques for seismic performance assessment of infrastructure systems. Eng. Struct. 250, 112883.
Meghan, G., Maya, W., 2021. Global Energy Demand to Grow 47% by 2050, with Oil Still Top Source. US EIA.
Meng, F., Wong, L.N.Y., Zhou, H., 2021a. Rock brittleness indices and their applications to different fields of rock engineering: a review. J. Rock Mech. Geotech. Eng. 13 (1), 221–247.
Meng, Y., Wang, Z., Zhang, L., He, C., Wen, R., Zhang, L., Wang, X., 2021b. Experimental evaluation on the conductivity of branch fracture with low sand laying concentration and its influencing factors in shale oil reservoirs. Lithosphere 2021.
Merzoug, A., Ellafi, A., 2023. Optimization of child well hydraulic fracturing design: a bakken case study. In: SPE Oklahoma City Oil and Gas Symposium/Production and Operations Symposium. SPE. D021S006R004.

Mgimba, M.M., Jiang, S., Mwakipunda, G.C., 2022. The identification of normal to underpressured formations in the Southeastern Sichuan basin. J. Petrol. Sci. Eng. 219, 111085.
Mgimba, M.M., Jiang, S., Ngole, W., 2023a. Optimization of hydraulic fracture treatment parameters for normally pressured Longmaxi and Wufeng shales in the southeastern Sichuan Basin in China. J. Energy Eng. 149 (2), 04023004.
Mgimba, M.M., Jiang, S., Nyakilla, E.E., Mwakipunda, G.C., 2023b. Application of GMDH to Predict Pore Pressure from Well Logs Data: A Case Study from Southeast Sichuan Basin, China. Natural Resources Research, pp. 1–21.
Mkono, C.N., Chuanbo, S., Mulashani, A.K., Mwakipunda, G.C., 2023. Deep Learning Integrated Approach for Hydrocarbon Source Rock Evaluation and Geochemical Indicators Prediction in the Jurassic-Paleogene of the Mandawa Basin, SE Tanzania. Energy, 129232.
Mulashani, A.K., Shen, C., Nkurlu, B.M., Mkono, C.N., Kawamala, M., 2022. Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data. Energy 239, 121915.
Munoz, H., Taheri, A., Chanda, E., 2016. Fracture energy-based brittleness index development and brittleness quantification by pre-peak strength parameters in rock uniaxial compression. Rock Mech. Rock Eng. 49, 4587–4606.
Mustafa, A., Tariq, Z., Abdulraheem, A., Mahmoud, M., Kalam, S., Khan, R.A., 2022. Shale brittleness prediction using machine learning—a Middle East basin case study. AAPG (Am. Assoc. Pet. Geol.) Bull. 106 (11), 2275–2296.
Nazar, S., Yang, J., Wang, X.-E., Khan, K., Amin, M.N., Javed, M.F., Althoey, F., Ali, M., 2023. Estimation of strength, rheological parameters, and impact of raw constituents of alkali-activated mortar using machine learning and SHapely Additive exPlanations (SHAP). Construct. Build. Mater. 377, 131014.
Niaki, M.H., Ahangari, M.G., Izadi, M., Pashaian, M., 2023. Evaluation of fracture toughness properties of polymer concrete composite using deep learning approach. Fatig. Fract. Eng. Mater. Struct. 46 (2), 603–615.
Nie, H., Chen, Q., Zhang, G., Sun, C., Wang, P., Lu, Z., 2021. An overview of the characteristic of typical Wufeng–Longmaxi shale gas fields in the Sichuan Basin, China. Nat. Gas. Ind. B 8 (3), 217–230.
Nouri, M., Khanlari, G., Rafiei, B., Sarfarazi, V., Zaheri, M., 2022. Estimation of brittleness indexes from petrographic characteristics of different sandstone types (cenozoic and mesozoic sandstones), Markazi Province, Iran. Rock Mech. Rock Eng. 55 (4), 1955–1995.
Omotehinwa, T.O., Oyewola, D.O., Dada, E.G., 2023. A light gradient-boosting machine algorithm with tree-structured parzen estimator for breast cancer diagnosis. Healthc. Anal., 100218
Ore, T., Gao, D., 2023. Prediction of reservoir brittleness from geophysical logs using machine learning algorithms. Comput. Geosci. 171, 105266.
Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F., El-Shafie, A., 2021. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng. J. 12 (2), 1545–1556.
Peng, J., Xu, C., Dai, B., Sun, L., Feng, J., Huang, Q., 2022. Numerical investigation of brittleness effect on strength and microcracking behavior of crystalline rock. Int. J. GeoMech. 22 (10), 04022178.
Qun, L., Yun, X., Bo, C., Baoshan, G., Xin, W., et al., 2022. Progress and prospects of horizontal well fracturing technology for shale oil and gas reservoirs. Petrol. Explor. Dev. 49 (1), 191–199.
Rickman, R., Mullen, M., Petre, E., Grieser, B., Kundert, D., 2008. A practical use of shale petrophysics for stimulation design optimization: all shale plays are not clones of the Barnett Shale. SPE Annu. Tech. Conf. Exhib.? SPE 144687, 1–11. SPE-115258-MS. https://doi.org/10.2118/115258-MS.
Rigatti, S.J., 2017. Random forest. J. Insur. Med. 47 (1), 31–39.
Rybacki, E., Meier, T., Dresen, G., 2016. What controls the mechanical properties of shale rocks?–Part II: brittleness. J. Petrol. Sci. Eng. 144, 39–58.
Rybacki, E., Reinicke, A., Meier, T., Makasi, M., Dresen, G., 2015. What controls the mechanical properties of shale rocks?–Part I: strength and Young's modulus. J. Petrol. Sci. Eng. 135, 702–722.
Ryu, B., Wang, L., Pu, H., Chan, M.K., Chen, J., 2022. Understanding, discovery, and synthesis of 2D materials enabled by machine learning. Chem. Soc. Rev. 51 (6), 1899–1925.
Saha, S., Gayen, A., Gogoi, P., Kundu, B., Paul, G.C., Pradhan, B., 2022. Proposing novel ensemble approach of particle swarm optimized and machine learning algorithms for drought vulnerability mapping in Jharkhand, India. Geocarto Int. 37 (25), 8004–8035.
Sarvi, F., Heuss, M., Aliannejadi, M., Schelter, S., de Rijke, M., 2022. Understanding and mitigating the effect of outliers in fair ranking. Proc. Fifteenth ACM Int. Conf. Web Search and Data Mining 861–869.
Shalaeva, D., Kukartseva, O., Tynchenko, V., Kukartsev, V., Aponasenko, S., Stepanova, E., 2020. Analysis of the development of global energy production and consumption by fuel type in various regions of the world. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, 012025.
Shi, C., Zhang, F., 2023. A forest fire susceptibility modeling approach based on integration machine learning algorithm. Forests 14 (7), 1506.
Shi, X., Liu, G., Cheng, Y., Yang, L., Jiang, H., Chen, L., Jiang, S., Wang, J., 2016a. Brittleness index prediction in shale gas reservoirs based on efficient network models. J. Nat. Gas Sci. Eng. 35, 673–685.
Shi, X., Liu, G., Jiang, S., Chen, L., Yang, L., 2016b. Brittleness Index Prediction from Conventional Well Logs in Unconventional Reservoirs Using Artificial Intelligence, International Petroleum Technology Conference. IPTC. D021S033R004.
Shi, X., Wang, J., Ge, X., Han, Z., Qu, G., Jiang, S., 2017. A new method for rock brittleness evaluation in tight oil formation from conventional logs and petrophysical data. J. Petrol. Sci. Eng. 151, 169–182.
Shi, X., Wang, M., Wang, Z., Wang, Y., Lu, S., Tian, W., 2021. A brittleness index evaluation method for weak-brittle rock by acoustic emission technique. J. Nat. Gas Sci. Eng. 95, 104160.
Song, Z., Zhang, J., Zhang, Y., Dong, X., Wang, S., 2023. Characterization and evaluation of brittleness of deep bedded sandstone from the perspective of the whole life-cycle evolution process. Int. J. Min. Sci. Technol. 33 (4), 481–502.
Sun, D., Lonbani, M., Askarian, B., Jahed Armaghani, D., Tarinejad, R., Thai Pham, B., Huynh, V.V., 2020. Investigating the applications of machine learning techniques to predict the rock brittleness index. Appl. Sci. 10 (5), 1691.
Sun, L., Koopialipoor, M., Jahed Armaghani, D., Tarinejad, R., Tahir, M., 2021. Applying a meta-heuristic algorithm to predict and optimize compressive strength of concrete samples. Eng. Comput. 37, 1133–1145.
Uddin, S., Haque, I., Lu, H., Moni, M.A., Gide, E., 2022. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci. Rep. 12 (1), 6256.
Wang, A.X., Chukova, S.S., Nguyen, B.P., 2023a. Ensemble k-nearest neighbors based on centroid displacement. Inf. Sci. 629, 313–323.
Wang, J., Tan, X., Wang, J., Zhang, H., Zhang, Y., et al., 2021. Characteristics and genetic mechanisms of normal-pressure fractured shale reservoirs: a case study from the Wufeng–Longmaxi formation in southeastern Chongqing, China. Front. Earth Sci. 9, 661706.
Wang, M., Zhao, G., Liang, W., Wang, N., 2023b. A comparative study on the development of hybrid SSA-RF and PSO-RF models for predicting the uniaxial compressive strength of rocks. Case Stud. Constr. Mater., e02191
Wang, Y., Xie, F., Zhao, T., Li, Z., Li, M., Liu, D., 2022. IGBT Status Prediction Based on PSO-RF with Time-Frequency Domain Features, 2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, pp. 337–341.
Wood, D.A., 2021. Brittleness index predictions from Lower Barnett Shale well-log data applying an optimized data matching algorithm at various sampling densities. Geosci. Front. 12 (6), 101087.
Xi, B., He, J., Li, H., 2023. Integration of machine learning models and metaheuristic algorithms for predicting compressive strength of waste granite powder concrete. Mater. Today Commun., 106403
Xia, Y., Li, L., Tang, C., Li, X., Ma, S., Li, M., 2017. A new method to evaluate rock mass brittleness based on stress–strain curves of class I. Rock Mech. Rock Eng. 50, 1123–1139.
Xia, Y., Zhou, H., Zhang, C., He, S., Gao, Y., Wang, P., 2022. The evaluation of rock brittleness and its application: a review study. Eur. J. Environ. Civil Eng. 26 (1), 239–279.
Xie, H., Lu, J., Li, C., Li, M., Gao, M., 2022. Experimental study on the mechanical and failure behaviors of deep rock subjected to true triaxial stress: a review. Int. J. Min. Sci. Technol. 32 (5), 915–950.
Xu, R., Li, Z., Jin, Y., 2022. Brittleness effect on acoustic emission characteristics of rocks based on a new brittleness evaluation index. Int. J. GeoMech. 22 (10), 04022185.
Yang, C., Chen, M., Yuan, Q., 2021. The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: an exploratory analysis. Accid. Anal. Prev. 158, 106153.
Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415, 295–316.
Yaro, A.S., Maly, F., Prazak, P., 2023. Outlier detection in time-series receive signal strength observation using Z-score method with S n scale estimator for indoor localization. Appl. Sci. 13 (6), 3900.
Ye, Y., Tang, S., Xi, Z., Jiang, D., Duan, Y., 2022. A new method to predict brittleness index for shale gas reservoirs: insights from well logging data. J. Petrol. Sci. Eng. 208, 109431.
Zeng, Q., Chen, S., He, P., Yang, Q., Guo, X., et al., 2018. Quantitative seismic prediction of shale gas sweet spots in lower silurian Longmaxi formation, weiyuan area, Sichuan Basin, SW China. Petrol. Explor. Dev. 45 (3), 406–414.
Zhang, D., Ranjith, P., Perera, M., 2016. The brittleness indices used in rock mechanics and their application in shale hydraulic fracturing: a review. J. Petrol. Sci. Eng. 143, 158–170.
Zhang, F., Deng, S., Zhao, H., Liu, X., 2022. A new hybrid method based on sparrow search algorithm optimized extreme learning machine for brittleness evaluation. J. Appl. Geophys. 207, 104845.
Zhang, J., Ma, X., Zhang, J., Sun, D., Zhou, X., Mi, C., Wen, H., 2023. Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model. J. Environ. Manag. 332, 117357.
Zhou, M., Lin, F., Hu, Q., Tang, Z., Jin, C., 2020. AI-enabled diagnosis of spontaneous rupture of ovarian endometriomas: a PSO enhanced random forest approach. IEEE Access 8, 132253–132264.