ACCSCIENCE PUBLISHING

Journal of Seismic Exploration

ARTICLE

Predicting channel sandstone thickness through a VIF-NRBO-XGBoost model

Weichao Zhang[1], Junhua Zhang[1]*, Zheng Huang[1], Jingqiang Yu[2], Deyong Feng[2], and Shugang Wang[2]

[1]School of Geosciences, China University of Petroleum (East China), Qingdao, Shangdong, China

[2]Geophysical Research Institute, Shengli Oilfield Company, SINOPEC, Dongying, Shangdong, China

Abstract

High-precision sand thickness data are fundamentally important for optimizing exploration strategies in petroleum geology. In the Chengbei work area of the Jiyang Depression, the stratigraphic channels are chaotically developed, with channels of varying sizes in different strata overlapping, intersecting, and exhibiting narrow widths. The actual well-seismic relationship is poor. Therefore, individual seismic attributes in this area exhibit extremely low correlation with channel sandstone thickness. Conventional attributes such as root mean square amplitude show no distinct channel characteristics, necessitating the integration of multiple seismic attributes for effective prediction. Moreover, the high multicollinearity among seismic attributes introduces significant interference in prediction results. Therefore, this study integrates the Pearson correlation coefficient and variance inflation factor (VIF) to optimize seismic attribute selection, effectively eliminating redundant attributes and those with low correlation. To further enhance prediction accuracy and address the significant bias inherent in single-model predictions, this study introduces the ensemble learning XGBoost model, which integrates predictions from multiple weak learners to improve the precision of sandstone thickness estimate. The Newton–Raphson-based optimization algorithm was employed to fine-tune the XGBoost parameters. Results from test wells demonstrate a remarkable improvement in prediction accuracy, achieving reliable sandstone thickness estimation despite poor well-seismic correlations. This research provides valuable insights and offers a widely applicable methodology for predicting the thickness of complex channel sand bodies.

Keywords: River channel sand body; Thickness prediction; Variance inflation factor; Newton-Raphson based optimization optimization; XGBoost

***Correspondence author:**
Junhua Zhang
(zjh@upc.edu.cn)

Citation: Zhang W, Zhang J, Huang Z, Yu J, Feng D, Wang S. Predicting channel sandstone thickness via a VIF-NRBO-XGBoost model. *J Seismic Explor*. 2025;34(3):1-12.
doi: 10.36922/JSE025290037

Received: July 18, 2025

Revised: August 27, 2025

Accepted: September 2, 2025

Published online: September 22, 2025

1. Introduction

Reservoir characterization constitutes a critical component in oil and gas field exploration and development. Scholars in related fields have conducted innovative research on reservoir thickness prediction, enhanced wettability characterization accuracy, and sandstone reservoir petrophysical properties.[1-3] Accurate prediction of

reservoir thickness is fundamental to detailed reservoir characterization and optimal exploration well placement, with increasingly stringent requirements for prediction-match rates. Given the high costs associated with acquiring fundamental seismic data, fully leveraging seismic data for reservoir thickness prediction holds significant importance for cost reduction and efficiency improvement in hydrocarbon exploration. To better utilize seismic data, geophysicists specializing in seismic data processing have integrated emerging technologies such as deep learning networks with wavelet transform methods to enhance seismic data resolution.[4] Seismic attributes, which are key information extracted from seismic data, contain abundant reservoir characteristics. Channel sand bodies represent one of the most important reservoir types in continental petroliferous basins. The Chengbei work area of Jiyang Depression studied in this paper exhibits chaotic channel development, where channel sand bodies demonstrate poor well-seismic relationships due to unfavorable conditions, including thin individual layers, narrow channel widths, severe overlapping and intersecting patterns, and multiple interbedded layers. These factors result in weak correlations between individual seismic attributes and sand body characteristics, necessitating multi-attribute seismic prediction. However, the strong multicollinearity among different seismic attributes precludes the simple superposition of multiple attributes with relatively strong correlations to sand body features for thickness prediction.[5]

The reservoir prediction for such complex channel sand bodies in this area has become a challenging issue, urgently requiring a novel method capable of effectively predicting sandstone thickness in such contexts.

In the field of reservoir thickness prediction, numerous studies have been conducted by petroleum geophysicists. Widess[6] first proposed estimating thin-bed thickness using reflection amplitude, but this method was only applicable to ideal reservoirs with equal-magnitude and opposite-polarity reflection coefficients. Chung and Lawton[7] improved upon this approach, achieving some enhancement in the prediction accuracy for very thin layers. However, the amplitude values remained constrained by the absolute values of the top and bottom reflection coefficients of the sand bodies, resulting in poor performance with actual data. Multi-attribute inversion has also been employed for sand body thickness prediction, utilizing seismic attributes sensitive to sand thickness combined with nonlinear optimization algorithms to calculate thickness. Nevertheless, this method suffers from low computational efficiency and is only effective in well-controlled areas, performing poorly in non-well-controlled regions.[8,9] Some scholars have proposed spectral decomposition techniques, using the "spectral notch" period to determine thin-bed thickness.[10-12] However, the "spectral notch" phenomenon is significantly influenced by factors such as wavelet bandwidth, limiting its practical application. Other approaches include identifying channel boundaries and predicting sand thickness using peak frequency-to-amplitude ratios, but these methods require high well-seismic correlation and are unsuitable for complex channel sand bodies with poor well-seismic relationships.[13] Barnes *et al.*[14] analyzed the relationship between frequency and reservoir thickness, establishing a corresponding formula for thickness distribution. However, this method shows low accuracy in complex areas with overlapping channels. Wang *et al.*[15] applied supervised learning based on fully connected neural networks to establish a nonlinear mapping between wavelet time-frequency components of seismic data and reservoir sand thickness, which, to some extent, reduced errors in validation wells.

Modern regression analysis frequently employs machine learning implementations, particularly tree-based ensemble methods like Random Forest, and kernel transformation techniques such as support vector regression (SVR) have demonstrated promising results in predicting sand body thickness.[16,17] While SVR models offer advantages for small-sample predictions and are theoretically suitable for areas with limited well data, their reliance on kernel functions for spatial mapping limits their ability to accurately handle nonlinear problems, resulting in weak nonlinear modeling capabilities. Chopra and Marfurt[18] were the first to utilize supervised learning algorithms, such as neural networks, to map multiple preferred attributes into reservoir thickness. Some researchers have employed eXtreme Gradient Boosting (XGBoost) models for sand thickness prediction, achieving favorable outcomes.[19] Furthermore, the XGBoost algorithm has found extensive utilization across multiple domains such as transportation, medicine, environment, and computer science.[20-23] Liu *et al.*[24] employed spectral decomposition-derived seismic characteristics combined with stacked generalization methodology to estimate reservoir thickness, which improved accuracy compared to other models. Currently, among various machine learning approaches, ensemble learning models show the most significant performance. However, challenges remain in optimal seismic attribute selection and parameter optimization for these ensemble models.

Based on the above research background, this paper proposes a VIF-NRBO-XGBoost reservoir thickness prediction model. To address the issues of strong multicollinearity among seismic attributes and low correlation between individual seismic attributes and reservoir thickness in complex channel sand bodies, this study combines variance inflation factor (VIF) and

Pearson correlation coefficient to conduct multicollinearity analysis and optimal seismic attribute selection.[25] To overcome the large prediction errors of single models and further improve prediction accuracy, an ensemble learning XGBoost model is introduced to enhance sand body thickness prediction precision by integrating predictions from multiple weak learners.[26] VIF serves as a diagnostic tool for detecting multicollinearity in multiple linear regression models, effectively eliminating redundant seismic attribute information. In general, a VIF value exceeding the threshold of 10 indicates unacceptable strong multicollinearity. Tree-based XGBoost ensemble learning demonstrates superior predictive performance for poor-quality data. However, this algorithm involves numerous parameters whose default settings typically fail to maximize model performance. Manual parameter adjustment proves excessively laborious and blind, making it practically infeasible. Currently, common parameter optimization methods include particle swarm optimization (PSO) and Bayesian optimization algorithms. For sand thickness prediction, PSO performs relatively poorly due to limited well data samples. Although Bayesian optimization shows improvement over PSO, it tends to converge to local optima, making it still challenging to find optimal parameter combinations for channel sand bodies with inherently poor well-seismic relationships. This study employs the Newton–Raphson-based optimization (NRBO) for model hyperparameter optimization.[27] The algorithm utilizes the Newton–Raphson search rule (NRSR) and the Trap Avoidance Operator (TAO) mechanisms to explore the search domain and enhance convergence speed. NRBO exhibits strong evolutionary capability, fast search speed, and excellent optimization performance. Finally, the prediction results are compared with other models to demonstrate the reliability of the proposed method.

2. Methodology

2.1. Variance inflation factor

Multicollinearity refers to the existence of linear relationships among independent variables. The VIF is a metric used to quantify the severity of multicollinearity among features in a regression model. A higher VIF value indicates stronger multicollinearity between the features. The VIF is calculated using the following formula:

$$VIF = \frac{1}{1-R_i^2} \tag{I}$$

Where R_i^2 represents the determination coefficient quantifying the linear relationship between the i-th selected feature and other features in the dataset. The computational method sequentially designates each feature as the response variable while considering the remaining features as predictors, fitting a regression model accordingly, and finally computes the ratio of mean squared errors between the independent and dependent variables. A VIF value near 1 suggests that the feature exhibits negligible multicollinearity. In general, two thresholds are set: when 5 < VIF < 10, it indicates relatively severe multicollinearity for that feature, requiring careful consideration; when VIF ≥ 10, it signifies extremely strong multicollinearity, necessitating elimination.

2.2. Fundamental principles of the XGBoost model

XGBoost represents an enhanced machine learning framework derived from the gradient boosting decision tree (GBDT) architecture, constituting an advanced implementation within the gradient boosting algorithmic paradigm. It consists of multiple decision trees that combine predictions from several weak learners to produce the final predictive outcome. Reservoir thickness prediction represents a typical regression problem, generally expressed through the following regression prediction formula:

$$\hat{y}_c = \sum_{k=1}^{K} f_k\left(x_c\right) \tag{II}$$

Where x_c represents the input sample, $f_k\ (x_c)$ is the prediction result calculated by the k-th tree, and by applying the principle of ensemble learning, the prediction results of the k trees are superimposed to obtain the final prediction result $\hat{y}_c$ of XGBoost. XGBoost assigns weights to each tree, and the subsequent trees will focus on the prediction information from the previous trees. Through multiple rounds of iterations, they converge to the final prediction result. Moreover, a regularization term is added to increase the model complexity:

$$O = \sum_{c=1}^{n} L\left(y_c, \hat{y}_c\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{III}$$

Where O represents the objective function established, L is the loss function to be calculated, and Ω is the regularization term added. Different from the conventional GBDT methods, the regularization term of XGBoost is:

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \sum_{q=1}^{T} \omega_q^2 \tag{IV}$$

Where γ represents the penalty factor, T indicates the number of leaf nodes, λ is the regularization parameter for leaf weights, ω_q represents the weight assigned to the leaf node at this time, and the regularization term is used to prevent the decision tree from being too large in scale, limit the number of leaf nodes, improve the model's out-of-sample performance, and mitigate overfitting risks through regularization constraints. The loss function is:

$$l(y_c, \hat{y}_c) = (y_c - \hat{y}_c)^2 \quad \text{(V)}$$

The XGBoost algorithm constructs its optimization objective function by integrating the prediction error term from the tree ensemble model with the model complexity regularization constraints:

$$W^{(U)} = \sum_{C=1}^{n} J\left(y_C, \hat{y}_C^{(t-1)} + f_t(x_C)\right) + \Omega\left(f_U\right) + C \quad \text{(VI)}$$

Where C represents a constant, and the target function is expanded using the Taylor series:

$$W^{(\ )} = \sum \left\{ J\left(y_C\ \hat{y}_C^{(\ )}\right) + g_C f_U\left(x_c\right) + -h_C f_U^2\left(x_C\right) \right\}$$
$$+\gamma T + -\lambda \sum \ \omega\ + C \quad \text{(VII)}$$

Where $g_c = \partial_{\hat{y}^{(U-1)C}} l(y_C, \hat{y}_C^{(U-1)})$, $h_c = \partial^2_{\hat{y}^{(U-1)}} l(y_C, \hat{y}_C^{(U-1)})$, denote the initial and successive rate-of-change measures in the differentiation hierarchy of the prediction error with respect to the model. Taking the first-order derivative of ω_q, we obtain the optimal objective function of XGBoost:

$$W = -\frac{1}{2}\sum_{q=1}^{T} \frac{G2_q}{H_q + \lambda} + \lambda T \quad \text{(VIII)}$$

The formula provides a structural scoring mechanism for tree models, with lower numerical values indicating superior topological configurations.

Take the derivative of **Equation VI** to obtain the optimal solution as follows:

$$\omega_q^* = -\frac{F_q}{R_q + \lambda} \quad \text{(IX)}$$

Where $F_q = \sum_{c \in I_q} g_c$, $R_q = \sum_{c \in I_q} h_c$ represent the sum of the first-order derivatives and the sum of the second-order derivatives of all input data mapped to leaf node q. I_q is the sample set of leaf nodes.

2.3. The principle of NRBO method

The NRBO is a novel metaheuristic optimization method whose inspiration primarily stems from two key principles: The NRSR and the TAO. By employing NRSR and TAO, the algorithm explores the search domain while enhancing convergence speed. NRBO exhibits strong evolutionary capabilities, rapid search performance, and excellent optimization ability.

(1) Exploratory starting point configuration: Throughout the primary population establishment process, NRBO creates a uniformly distributed candidate population spanning the solution space boundaries, which serves as the foundation for subsequent iterative refinement. Suppose there are N populations; NRBO uses Equation 10 to generate the random population:

$$h_o^k = lb + rand \times \left(ub - lb\right), k = 1,2,\ldots,N_{op}, p = 1,2,\ldots,\dim \quad \text{(X)}$$

In the population matrix representation, element h_o^k stores the position value of the k-th candidate solution in its p-th dimensional component, r and represents a random number within the range of (0, 1), and the search space is constrained by lb (minimum value) and ub (maximum value) for each parameter. Formula 11 depicts the population matrix of all dimensions:

$$H_k = \begin{bmatrix} h_1^1 & h_2^1 & \cdots & h_{\dim}^1 \\ h_1^2 & h_2^2 & \cdots & h_{\dim}^2 \\ \vdots & \vdots & \ddots & \vdots \\ h_1^{N_{oP}} & h_2^{N_{OP}} & \cdots & h_{\dim}^{N_{OP}} \end{bmatrix}_{N_{0P} \times \dim} \quad \text{(XI)}$$

(2) The NRSR is developed by adapting the classical Newton–Raphson method, with dual objectives of enhancing trend discovery capability and improving convergence rate. The Newton method is an iterative process used to find the roots of an equation. It obtains the next estimate by performing a two-dimensional Taylor Series (TS) around the current estimated minimum value. The iteration continues until the first derivative of the function approaches the threshold, and the minimum point estimate is finally determined. **Formula XII** represents the second-order Taylor Series of $v(h)$ at h_0:

$$v(h) = \frac{1}{0!} f(h_0) + \frac{1}{1!}(h - h_0) v'(h_0) + \frac{1}{2!}(h - h_0)^2 v'\left(h_0\right) \quad \text{(XII)}$$

By taking the derivative of both sides of the above equation and setting it equal to zero, we obtain the following equation:

$$v'\left(h\right) = v'\left(h_0\right) + v''\left(h_0\right)\left(h - h_0\right) = 0 \quad \text{(XIII)}$$

The above equation can be solved as:

$$h = h_0 - \frac{v'\left(h_0\right)}{v''\left(h_0\right)} \quad \text{(XIV)}$$

The above process is repeated until a point with zero derivative is obtained. **Formula XV** is the iterative formula for the obtained point:

$$h_{n+1} = h_n - \frac{v'\left(h_n\right)}{fv''\left(h_n\right)} \quad \text{(XV)}$$

In order to obtain NRSR from the above equation, the second-order Taylor series of $v(h + \Delta h)$ and $f(h-\Delta h)$ are written as follows:

$$v\left(h+\Delta h\right)=v\left(h\right)+v^{'}\left(h_0\right)\Delta h+\frac{1}{2!}v^{''}\left(h_0\right)\Delta h^2 \quad \text{(XVI)}$$

$$v\left(h-\Delta h\right)=v\left(h\right)-v^{'}\left(h_0\right)\Delta h+\frac{1}{2!}v^{''}\left(h_0\right)\Delta h^2 \quad \text{(XVII)}$$

By subtracting or adding **Formulas XVI and XVII**, the expressions of $v'(h)$ and $v''(h)$ can be obtained:

$$v^{'}\left(h\right)=\frac{v\left(h+\Delta h\right)-v\left(h-\Delta h\right)}{2\Delta h} \quad \text{(XVIII)}$$

$$v^{''}\left(h\right)=\frac{v\left(h+\Delta h\right)+v\left(h-\Delta h\right)-2v\left(h\right)}{\Delta h^2} \quad \text{(XIX)}$$

Substitute **Formulas XVIII and XIX** into **Formula XV**, and the updated root positions are as follows:

$$h_{n+1}=h_n-\frac{\left(v\left(h_n+\Delta h\right)-v\left(h_n-\Delta h\right)\right)\times\Delta h}{2\times\left(v\left(h_n+\Delta h\right)+v\left(h_n-\Delta h\right)-2\times v\left(h_n\right)\right)} \quad \text{(XX)}$$

Where $h_n + \Delta h$ and $h_n - \Delta h$ respectively represent the positions of adjacent x's to each other, and NRSR is defined as follows:

$$NRSR=randn\times\frac{\left(H_w-H_b\right)\times\Delta h}{2\times\left(H_w+XH_b-2\times h_n\right)} \quad \text{(XXI)}$$

Where *randn* generates random scalars drawn from the standard normal distribution ($\mu = 0$, $\sigma^2 = 1$). H_w and H_b, respectively, denote the worst and best positions.

$$\Delta h=rand\left(1,\dim\right)\times\left|H_b-H_n^{IT}\right| \quad \text{(XXII)}$$

Where H_b represents the current optimal solution, and $rand(1,dim)$ is a set of random numbers with *dim* decision variables. Then, by using NRSR, **Formula XI** is modified to:

$$h_{n+1}=h_n-NRSR \quad \text{(XXIII)}$$

A guidance parameter ρ is introduced to direct the population's positional updates toward the optimal solution region:

$$\rho=a\times\left(H_b-H_n^{IT}\right)+b\times\left(H_{s_1}^{IT}-H_{s_2}^{IT}\right) \quad \text{(XXIV)}$$

Where a and b are random numbers within the range of (0, 1), s_1 and s_2 are different integers selected, and the current position of the vector is updated by **Formula XXV**:

$$H1_n^{IT}=h_n^{IT}-\left(randn\times\frac{\left(H_w-H_b\right)\times\Delta h}{2\left(H_w+H_b-2\times H_n\right)}\right)+\left(a\left(H_b-H_n^{IT}\right)\right)+b\times\left(H_{s_1}^{IT}-H_{s_2}^{IT}\right) \quad \text{(XXV)}$$

Where the vector $H1_k^{IT}$ represents the updated position derived from h_k^{IT} through the enhanced NRSR, which constitutes an optimized variant of the standard Newton-Raphson Method (NRM). **Formula XXI** becomes:

$$NRSR=randn\times\frac{\left(y_w-y_b\right)\times\Delta h}{2\times\left(y_w+y_b-2\times h_n\right)} \quad \text{(XXVI)}$$

$$y_w=s_1\times\left(Mean\left(M_{K+1}+h_n\right)+s_1\times\Delta h\right) \quad \text{(XXVII)}$$

$$y_b=s_1\times\left(Mean\left(M_{K+1}+h_n\right)-s_1\times\Delta h\right) \quad \text{(XXVIII)}$$

$$M_{K+1}=h_K-randn\times\frac{\left(H_w-H_b\right)\times\Delta h}{2\times\left(H_w+H_b-2\times h_k\right)} \quad \text{(XXIX)}$$

Where y_w and y_b denote position vectors derived from M_{n+1} and h_k, respectively, where $s_1 \sim U(0,1)$, represents a uniformly distributed random coefficient. The candidate solution for the subsequent generation is determined by:

$$H_k^{IT}=h_k^{IT}-\left[randn\frac{\left(y_w-y_b\right)\Delta h}{2\left(h\left(y_w+y_b-2h_k\right)\right)}\right]+\left(a\left(H_b-H_k^{IT}\right)+b\left(H_k^{IT}-H_k^{IT}\right)\right) \quad \text{(XXX)}$$

(3) TAO: The TAO framework integrates an advanced optimization operator developed by Ahmadianfar *et al.*,[28] which significantly boosts NRBO's performance in real-world applications while mitigating local optimum convergence risks. This implementation activates when the stochastic variable rand (uniformly distributed in [0,1]) falls below the decisive factor DF (default threshold: 0.6). Then, the solution X_{TAO}^{IT} is generated using the following formula:

$$\begin{cases} X_{TAO}^{IT}=X_n^{IT+1}+\theta\times X(\mu_1\times x_n-\mu_2\times X_n^{IT}) \\ +\theta\times X\times\delta\times X(\mu_1\times Mean(X_n^{IT})-\mu_2\times X_n^{IT}),\mu_1<0.5 \\ X_{TAO}^{IT}=x_n+\theta\times X(\mu_1\times x_n-\mu_2\times X_n^{IT}) \\ +\theta\times X\times\delta\times X(\mu_1\times Mean(X_n^{IT})-\mu_2\times X_n^{IT}),\mu_1\geq 0.5 \end{cases} \quad \text{(XXXI)}$$

$$X_n^{IT+1}=X_{TAO}^{IT} \quad \text{(XXXII)}$$

Where *rand* is a random number, θ_1 and θ_2 are uniformly distributed random numbers within the range of (−1,1) and (−0.5,0.5), respectively. The parameters μ_1 and μ_2

are assigned stochastic values during initialization. The randomness of μ_1 and μ_2 prevents the population from falling into local optima.

2.4. VIF-NRBO-XGBoost reservoir prediction workflow

The VIF-NRBO-XGBoost-based prediction workflow for channel sand reservoir thickness proceeds as follows: First, seismic attributes are extracted and preliminarily optimized, prioritizing those with clear geological significance and superior quality. The selected seismic attributes then undergo outlier removal and normalization processing. Subsequently, VIF values are calculated for all extracted seismic attributes, combined with Pearson correlation coefficients for comprehensive attribute analysis. Hyperparameter selection for the XGBoost algorithm is accomplished through NRBO optimization. The processed seismic attributes are then paired with corresponding well-point thickness data to train the NRBO-XGBoost reservoir thickness prediction model. To ensure evaluation stability, K-fold cross-validation (with K = 5 in this study) is implemented, using the mean absolute error from five validation wells to assess prediction accuracy. Finally, the model predicts reservoir thickness for other target areas within the work zone. The complete workflow is illustrated in Figure 1.

3. Correlation analysis combined with VIF for selecting optimal seismic attributes

The typical lithofacies bodies in Jiyang Depression have rich reservoir types. The ancient river channel sand bodies are representative lithofacies among them. This paper takes Chengbei Oilfield as the research area, and the study section is the upper part of the Guantao Formation. The large and small river channels are superimposed and crossed, while the single sand body reservoir is thin. According to the geological meaning of seismic attributes and the comprehensive data quality, a total of 11 distinct seismic attributes from different categories were extracted from the target formation, including: Root mean square amplitude (RMS_amp), bandwidth (BW), zero-crossing count (ZCC), arc length (AL), energy half-time (EHT), average energy (AE), average instantaneous frequency (AIF), average amplitude (AA), positive amplitude sum (PAS), dominant frequency (DF), maximum amplitude (MA).

3.1. Correlation analysis of seismic attributes

In machine learning regression experiments, the Pearson correlation coefficient, scatter plots, and linear models are the three most commonly used methods. Figure 2 comprehensively displays the following: (i) Complete inter-variable linear dependencies are shown in the matrix upper triangle, quantifying how each of the 11 seismic attributes covaries with formation thickness at well locations; (ii) The lower triangle presents scatter plots of correlations between different attributes, as well as between all attributes and thickness, with overlaid linear regression lines. To better visualize the linear relationship between individual seismic attributes and thickness, along with statistical reliability, 95% confidence intervals are included in the scatter plots; (iii) The diagonal displays normalized distribution histograms and Kernel density estimation of the seismic attributes, clearly reflecting their distribution patterns. From the data and scatter plots as shown in Figure 2, individual seismic parameters demonstrate limited predictive capability for thickness estimation in reservoir formations, and the distribution of single attributes shows no significant patterns.

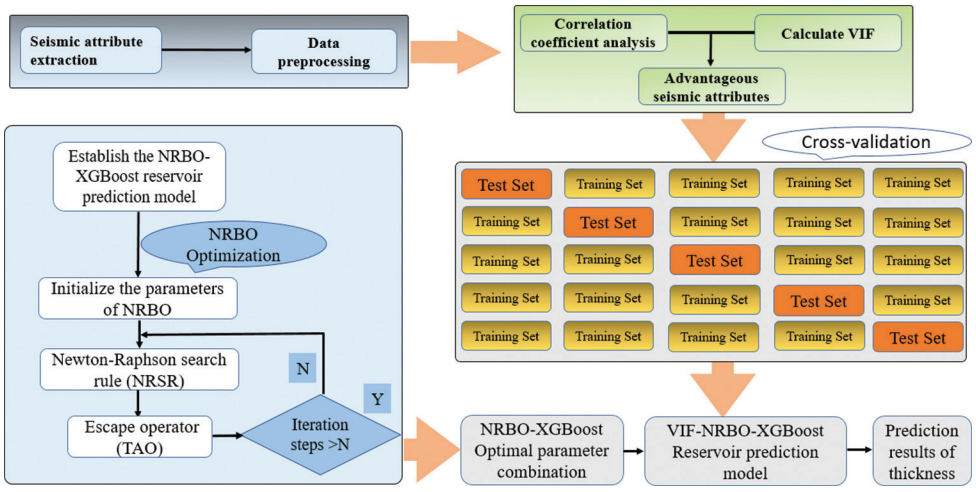

Figure 1. VIF-NRBO-XGBoost process for predicting reservoir thickness of riverbed sedimentary rocks
Abbreviations: NRBO: Newton-Raphson-based optimization; VIF: Variance inflation factor; XGBoost: eXtreme gradient boosting.

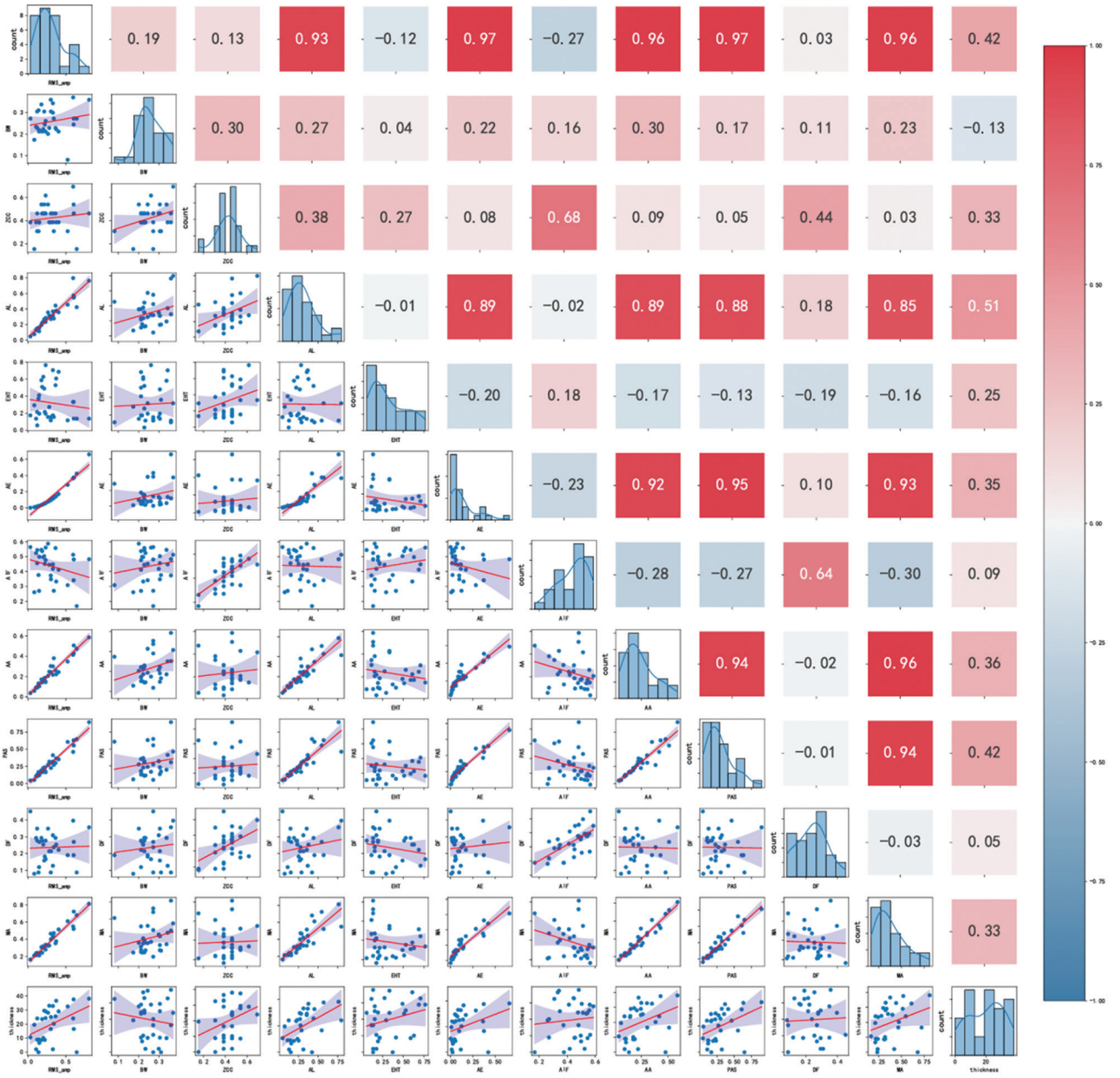

Figure 2. Correlation analysis

This further demonstrates the geological complexity of the study area.

3.2. Selection of VIF attributes

Before conducting attribute selection using VIF, to prevent the interference of seismic attributes with low correlation to reservoir thickness from affecting the attribute screening, leveraging the identified attribute-thickness correlations, the three seismic attributes with correlation <0.2 with reservoir thickness, namely bandwidth, AIF, and DF, were removed first.[29] Then, VIF analysis was conducted on the remaining seismic attributes. Figure 3 shows the VIF values and correlation coefficients of the remaining eight seismic attributes. It can be seen that the VIF value of the RMS amplitude is very high, indicating that there is severe multicollinearity between it and the other seismic attributes, and it must be eliminated. The VIF values of ZCC and EHT are very low, indicating that the multicollinearity of these two seismic attributes is very weak. In addition, the VIF values of AL, AE, AA, PAS, and MA are similar. As can be observed from Figure 2, among these four seismic attributes, AL shows the highest

correlation with thickness. Finally, three seismic attributes, namely AL, ZCC, and EHT, were retained for reservoir thickness prediction.

4. VIF-NRBO-XGBoost reservoir thickness prediction

To prevent overfitting or underfitting, considering the characteristics of limited sample data, the proportion of the test set is set to 15%. After multiple verifications, the general range of XGBoost's hyperparameters is found. Then, XGBoost is utilized to conduct prediction comparisons between the seismic attributes that have not undergone VIF screening and those that have undergone VIF screening. As shown in Figure 4, it can be observed that the degree of deviation of the prediction results of the seismic attributes after VIF screening is lower, and the prediction accuracy is higher.

The NRBO optimization, combined with cross-validation, is utilized to search for the optimal solution for the hyperparameters of XGBoost. The best parameter NRBO-

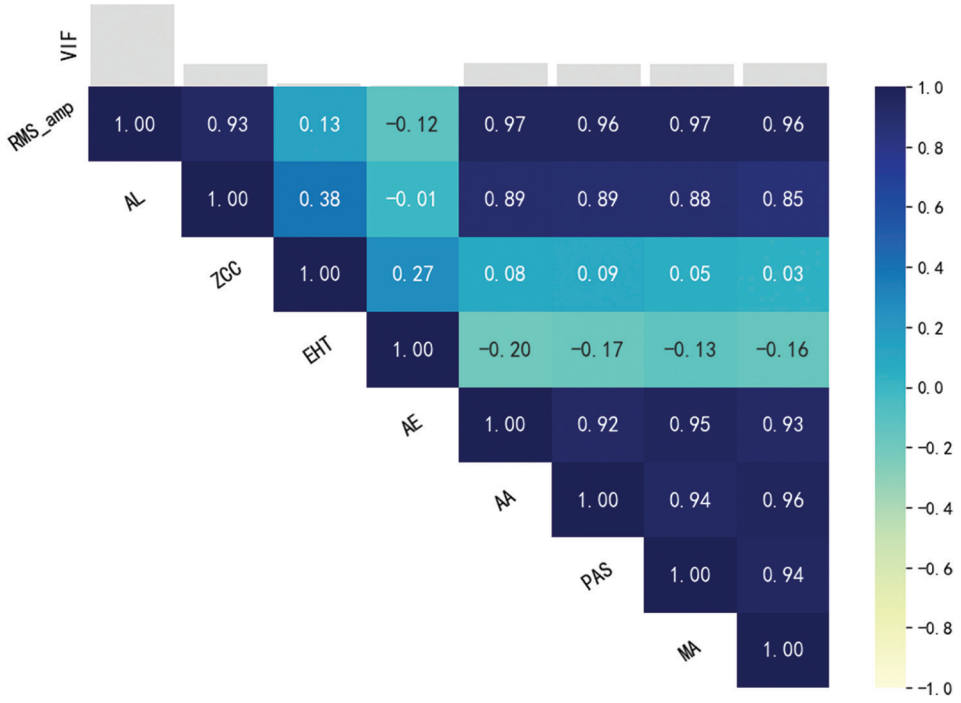

Figure 3. VIF of seismic attribute and correlation
Abbreviations: AA: Average amplitude; AE: Average energy; AL: Arc length; EHT: Energy half-time; MA: Maximum amplitude; PAS: Positive amplitude sum; RMS_amp: Root mean square amplitude; VIF: Variance inflation factor; ZCC: Zero-crossing count.

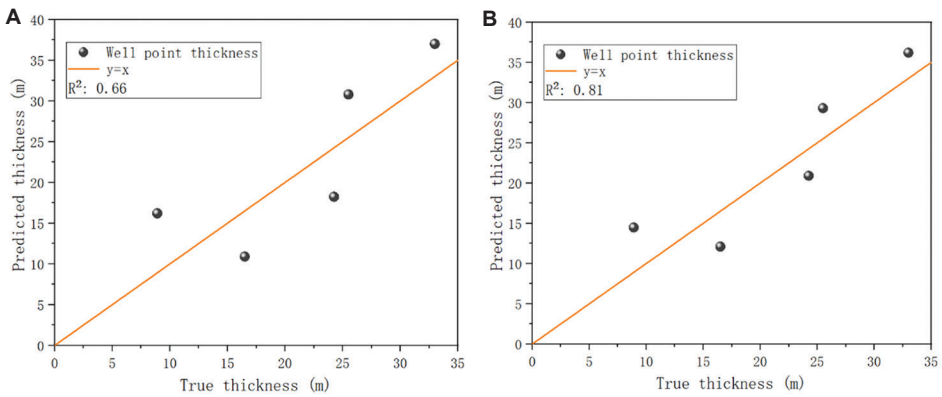

Figure 4. Comparison of XGBoost prediction results before (A) and after VIF screening (B)
Abbreviations: VIF: Variance inflation factor; XGBoost: eXtreme gradient boosting.

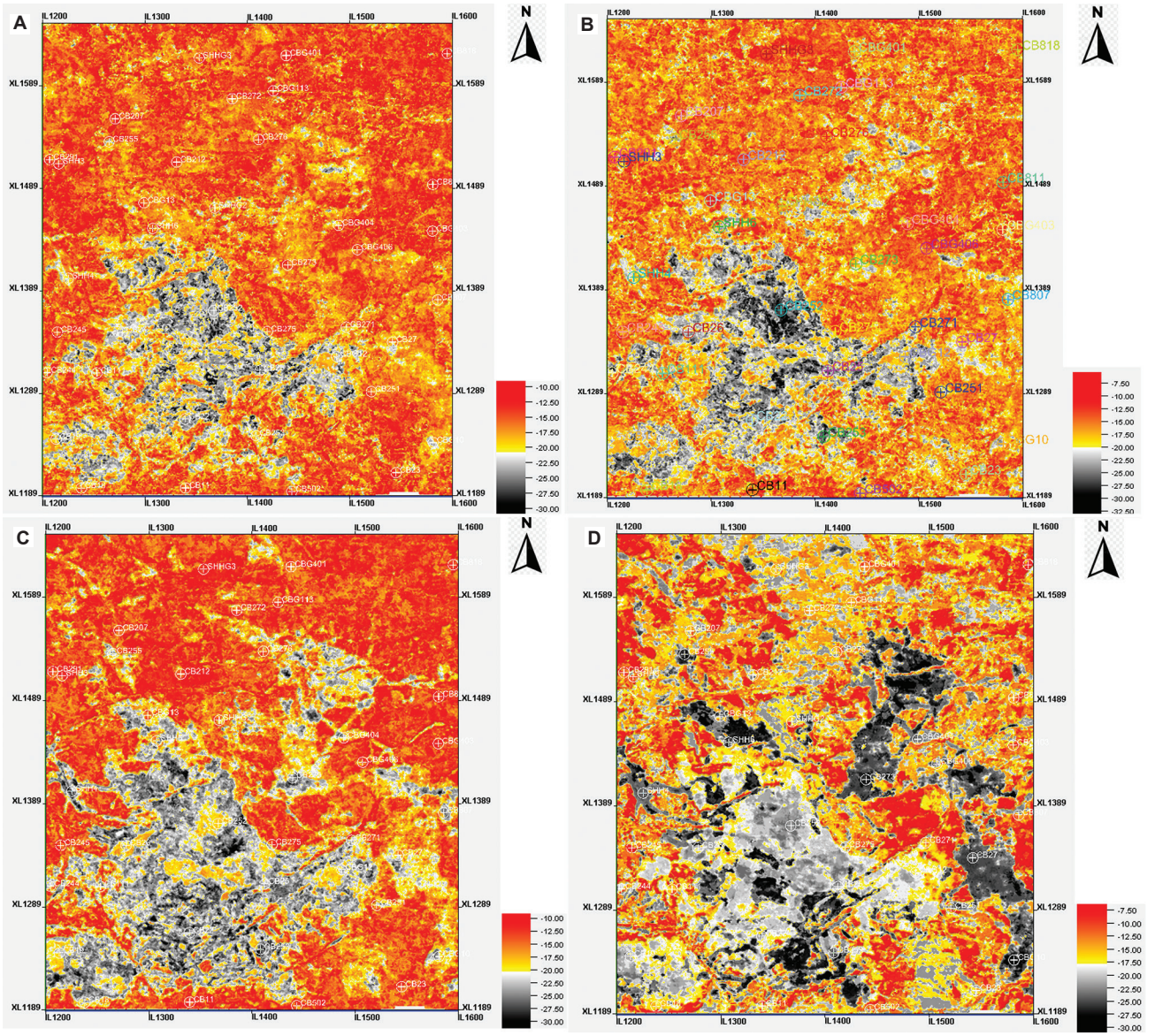

Figure 5. The prediction results of reservoir thickness of riverbed sand in the study area. (A) Predictive outputs from the SVM; (B) predictive outputs from the XGBoost, (C) predictive outputs from the VIF-XGBoost, and (D) predictive outputs from the VIF-NRBO-XGBoost.
Abbreviations: NRBO: Newton-raphson based optimization; SVM: Support vector machine; VIF: Variance inflation factor; XGBoost: eXtreme gradient boosting.

XGBoost is developed for sandstone thickness estimation in reservoir characterization, and the predictive outcomes are systematically benchmarked against conventional XGBoost results and Support Vector Machine (SVM) models that have not been optimized. The prediction results are shown in Figure 5, and the comparison of the average absolute error and R^2 of the prediction results of the four models for sand body thickness in the verification wells is presented in Table 1. Based on the prediction results, evidence suggests that the SVM model demonstrates low prediction accuracy with significant absolute errors, failing to capture the distinct morphological features of channel sand bodies. Although the VIF-XGBoost model provides a more accurate depiction of the eastern river channel sand bodies, its overall prediction accuracy remains inadequate. VIF-NRBO-XGBoost algorithm demonstrates dual capabilities in fluvial reservoir characterization, successfully capturing both the extensive channel systems in eastern sectors and accurately forecasting subtle channel deposits in southwestern regions.

The VIF-NRBO-XGBoost modeling results reveal distinct fluvial depositional patterns across the study

Table 1. Comparison of prediction results and mean absolute errors of four models for verification wells

Well name and evaluation metric	True thickness	SVM	XGBoost	VIF-XGBoost	VIF-NRBO-XGBoost
CB245	8.9 m	16.9 m	16.2 m	14.48 m	9.7 m
CB253	25.5 m	18.8 m	30.8 m	29.3 m	24.2 m
CB255	33 m	24.8 m	37 m	36.2 m	31.1 m
CB11	16.5 m	9.8 m	10.9 m	12.1 m	17.9 m
CB27	24.25 m	17.9 m	18.24 m	20.9 m	25.7 m
Mean absolute error	\	7.2 m	5.6 m	4.1 m	1.4 m
R^2	\	0.48	0.66	0.81	0.97

Abbreviations: NRBO: Newton–Raphson-based optimization; SVM: Support vector machine; VIF: Variance inflation factor; XGBoost: eXtreme gradient boosting.

area, with a prominent north-south-oriented channel belt dominating the eastern sector. Central regions exhibit maximum sandbody thickness accompanied by a gradual southeastward deflection of the channel axis. The southwestern domain contains smaller-scale channel features with potential tributary systems, displaying predominant northwest-to-southeast paleoflow orientations.

5. Discussion

To address the complex development of underground channel sand bodies in the Chengbei work area of the Jiyang Depression, characterized by chaotic, intersecting, and overlapping patterns, a novel VIF-NRBO-XGBoost model for sand body thickness prediction was introduced. The model was trained using 35 known wells and validated with five known wells (CB245, CB253, CB255, CB11, CB27), followed by a comprehensive prediction across the entire work area, effectively improving the thickness prediction accuracy for such complex channel sand bodies. The model primarily consists of the following steps:

First, 11 commonly used seismic attributes related to reservoir information were extracted and normalized. The Pearson correlation coefficient was employed to preliminarily screen these 11 seismic attributes, removing those with a correlation coefficient of <0.2 with sand body thickness. To prevent multicollinearity among the seismic attributes from affecting the prediction results, the remaining eight seismic attributes were subjected to multicollinearity analysis using VIF, and attributes with strong multicollinearity and redundant information were eliminated.

Table 2. Comparison of XGBoost model parameters before and after NRBO optimization

Model parameter	XGBoost	NRBO-XGBoost
n_estimators	150	193
max_depth	7	12
min_child_weight	3	1
learning_rate	0.04	0.059
colsample_bytree	0.5	0.57
gamma	6	4.5
alpha	3	3.559

Abbreviations: Alpha: Regularization coefficient; colsample_bytree: Feature random sampling ratio; gamma: Node splitting reduction coefficient; learning_rate: Learning rate; max_depth: Maximum tree depth; min_child_weight: Minimum leaf node weight; n_estimators: Number of decision trees; NRBO: Newton–Raphson-based optimization; XGBoost: eXtreme gradient boosting.

Due to the poor data quality in this region, single machine learning models exhibited significant prediction errors. An ensemble learning XGBoost model was introduced to enhance prediction accuracy by integrating the results of multiple weak learners. The performance of the XGBoost model largely depends on the selection of model parameters. In this study, the NRBO intelligent optimization algorithm was used to optimize the XGBoost model parameters, and the optimal parameter combination was employed for sand body thickness prediction, resulting in more refined channel sand body distribution predictions. Table 2 lists the seven core parameters of the XGBoost model before and after NRBO optimization: the number of decision trees (n_estimators), maximum tree depth (max_depth), minimum leaf node weight (min_child_weight), learning rate (learning_rate), feature random sampling ratio (colsample_bytree), node splitting reduction coefficient (gamma), and regularization coefficient (alpha).

Although the VIF-NRBO-XGBoost model outperforms other machine learning models in predicting the thickness of complex channel sand bodies with higher accuracy, the correlation analysis directly removed seismic attributes with extremely low correlation to thickness, potentially losing valuable information from these attributes. Future research will consider the modal information of seismic attributes to fully retain useful information from the discarded attributes. Additionally, further optimization of model parameters will be pursued to enhance the prediction accuracy of complex channel sand body thickness.

The data used in this study constitutes a small sample dataset. The performance of the aforementioned method on large sample datasets remains unclear and may require adjustments to the validation set ratio. The prediction accuracy of this method is somewhat dependent on data

quality and resolution, and the current model may exhibit uncertainties in predicting extremely thin sandstone layers. Future work will consider incorporating additional data sources, such as seismic attribute modalities, to further enhance the model's generalization capability.

6. Conclusion

This study proposes a novel sand body thickness prediction model—VIF-NRBO-XGBoost. The model utilizes multiple attributes for reservoir thickness prediction while fully considering the constraints of multicollinearity and correlation among seismic attributes, employing NRBO to optimize the parameters of the ensemble learning XGBoost model. Through its application in predicting complex channel sand bodies in the Chengbei area of Jiyang Depression, the reliability of the model was verified, with prediction results significantly outperforming other models. This will provide crucial support for detailed reservoir characterization and well placement in this region. The study not only offers new insights for reservoir thickness prediction in similar study areas, but also provides valuable references for predicting other reservoir parameters. It holds significant practical importance for hydrocarbon exploration and development.

Acknowledgments

None.

Funding

This study is supported by the National Science and Technology Major Project Management Office for New Oil and Gas Exploration and Development (NO. 2024ZD1400102); and Shengli Oilfield Jiyang Depression Typical Lithologic Body Project Fund (NO. 30200020-24-ZC0613-0062).

Conflict of interest

All authors declare no conflicts of interest.

Author contributions

Conceptualization: All authors
Formal analysis: All authors
Investigation: All authors
Methodology: All authors
Visualization: Weichao Zhang, Junhua Zhang
Writing–original draft: Weichao Zhang
Writing–review & editing: All authors

Availability of data

Data are available from the corresponding author on reasonable request. Note that certain datasets may not be shareable due to confidentiality reasons.

References

1. Wang S, Gu Z, Guo P, Zhao W. Comparative laboratory wettability study of sandstone, tuff, and shale using 12- MHz NMR T1- T2 fluid typing: Insight of shale. *SPE J*. 2024;29(9):4781-4803.

   doi: 10.2118/221496-PA

2. Ehsan M, Chen R, Manzoor U, *et al*. Unlocking thin sand potential: A data-driven approach to reservoir characterization and pore pressure mapping. *Geomech Geophys Geoenerg Georesour.* 2024;10:160.

   doi: 10.1007/s40948-024-00871-w

3. Liu L, Jin J, Liu J, *et al*. Mechanical properties of sandstone under *in-situ* high-temperature and confinement conditions. *Int J Miner Metall Mater*. 2025;32(4):778-787.

   doi: 10.1007/s12613-024-3047-9

4. Manzoor U, Ehsan M, Hussain M, Bashir Y. Improved reservoir characterization of thin beds by advanced deep learning approach. *Appl Comput Geosci*. 2024;23:100188.

   doi: 10.1016/j.acags.2024.100188

5. Dormann CF, Elith J, Bacher S, *et al*. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2012;36:27-46.

   doi: 10.1111/j.1600-0587.2012.07348.x

6. Widess MB. How thin is a thin bed? *Geophysics*. 1973;38(6):1176-1180.

   doi: 10.1190/1.1440403

7. Chung HM, Lawton DC. Amplitude responses of thin beds: Sinusoidal approximation versus Ricker approximation. *Geophysics*. 1995;60(1):19-307.

   doi: 10.1190/1.1443750

8. Long J. Neural network BP modeling of the relation between thin bed thickness and amplitude-frequency. *Oil Geophys Prospect*. 1995;30(6):817-822.

9. Zhuang D, Xiao C. Thin-bed thickness estimation using neural network. *Oil Geophys Prospect*. 1996;31(3):394-399.

10. Gridley JA, Partyka GA. *Processing and Interpretational Aspects of Spectral Decomposition: 67th Annual International Meeting of the Society of Exploration Geophysicists. Expanded Abstracts*; 1997. p. 1055-1058.

11. Partyka GA, Gridley JA, Lopez JA. Interpretational aspects of spectral decomposition in reservoir characterization. *Lead Edge*. 1999;18(3):353-360.

12. Ye T, Su J, Liu X. Application of seismic frequency division interpretation technology in predicting continental sandstone reservoir in the west of Sichuan province. *Geophys Prospect Petrol*. 2008;47(1):72-76.

13. Sun L, Zheng X, Shou H, Li J, Li Y. Quantitative prediction of channel sand bodies based on seismic peak attributes in the frequency domain and its application. *Appl Geophys*. 2010;7(1):10-17.

doi: 10.1007/s11770-010-0009-y

14. Barnes AE, Fink L, Laughlin K. *Improving Frequency Domain Thin Bed Analysis. SEG Technical Program Expanded Abstracts*. United States: SEG. 2004. p. 1929-1932.

doi: 10.1190/1.1851175

15. Wang Z, Gao D, Lei X, Wang D, Gao J. Machine learning-based seismic spectral attribute analysis to delineate a tight-sand reservoir in the Sulige gas field of central Ordos Basin, western China. *Mar Petrol Geol*. 2020;113:104136.

doi: 10.1016/j.marpetgeo.2019.104136

16. Li W, Yue D, Wang W, *et al*. sing multiple frequency-decomposed seismic attributes with machine learning for thickness prediction and sedimentary facies interpretation in fluvial reservoirs. *J Petrol Sci Eng*. 2019;177:1087-1102.

doi: 10.1016/j.petrol.2019.03.017

17. Yue D, Li W, Wang W, *et al*. Fused spectral-decomposition seismic attributes and forward seismic modelling to predict sand bodies in meandering fluvial reservoirs. *Mar Petrol Geol*. 2019;99:27-44.

doi: 10.1016/j.marpetgeo.2018.09.031

18. Chopra S, Marfurt KJ. Seismic attributes - A historical perspective. *Geophysics*. 2005;70(5):1SO-Z82.

doi: 10.1190/1.2098670

19. Liu S. A grain size profile prediction method based on combined model of extreme gradient boosting and artificial neural network and its application in sand control design. *SPE J*. 2024;29(6):2988-3002.

doi: 10.2118/219484-PA

20. Zhang D, Qian L, Mao B, *et al*. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Acess*. 2018;6:21020-21031.

doi: 10.1109/ACCESS.2018.2818678

21. Yu B, Qiu W, Chen C, *et al*. SubMito-XGBoost: Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics*. 2020;36(4):1074-1081.

doi: 10.1093/bioinformatics/btz734

22. Kounlavong K, Sadik L, Keawsawasvong S, Jamsawang P. Novel hybrid XGBoost-based soft computing models for predicting penetration resistance of buried pipelines in cohesive soils. *Ocean Eng*. 2024;311(2):118948.

doi: 10.1016/j.oceaneng.2024.118948

23. Liu X, Zhang J, Bai Q, *et al*. Research and application of sandstone thickness prediction method based on NRBO-XGBoost optimization method. *Comput Techn Geophys Geochem Explor*. 2024;46(2):146-153.

doi: 10.3969/j.issn.1001-1749.2024.02.03 (in Chinese)

24. Liu L, Li W, Du Y, *et al*. Reservoir prediction method of fusing frequency-decomposed seismic attributes using Stacking ensemble learning. *Oil Geophys Prospect*. 2024;59(1):12-22.

doi: 10.13810/j.cnki.issn.1000-7210.2024.01.002 (in Chinese)

25. Cheng J, Sun J, Yao K, Xu M, Cao Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim Acta A Mol Biomol Spectrosc*. 2022;268:120652.

doi: 10.1016/j.saa.2021.120652

26. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. USA; 2016. p. 785-794.

doi: 10.1145/2939672.2939785

27. Sowmya R, Premkumar M, Jangir P. Newton raphson-based optimizer: A new population-based metaheuristic algorithm for continuous optimization problems. *Eng Appl Artif Intell*. 2024;128(C):107532.

doi: 10.1016/j.engappai.2023.107532

28. Ahmadianfar I, Bozorg-Haddad O, Chu X. Gradient-based optimizer: A new metaheuristic optimization algorithm. Information *Sciences*. 2020;540:131-159.

doi: 10.1016/j.ins.2020.06.037

29. Lian Y, Luo J, Xue W, Zuo G, Zhang S. Cause-driven streamflow forecasting framework based on linear correlation reconstruction and long short-term memory. *Water Resour Manag*. 2022;36:1661-1678.

doi: 10.1007/s11269-022-03097-1