applied sciences

MDPI

Article

Prediction of Horizontal in Situ Stress in Shale Reservoirs Based on Machine Learning Models

Wenxuan Yu [1], Xizhe Li [1,2,3,*], Wei Guo [1], Hongming Zhan [1,*], Xuefeng Yang [4], Yongyang Liu [4], Xiangyang Pei [1], Weikang He [1,2,3], Longyi Wang [1,2,3] and Yaoqiang Lin [1]

[1] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing 100083, China; yuwxpetrochina@163.com (W.Y.); pkuguowei69@petrochina.com.cn (W.G.); peixy@petrochina.com.cn (X.P.); heweikang23@mails.ucas.ac.cn (W.H.); wly990914@163.com (L.W.); linyaoqiang990309@163.com (Y.L.)
[2] School of Engineering Science, University of the Chinese Academy of Sciences, Beijing 101400, China
[3] Institute of Porous Flow and Fluid Mechanics, Chinese Academy of Sciences, Langfang 065000, China
[4] Shale Gas Research Institute of Southwest Oil & Gas Field Branch, Chengdu 610051, China; yangxuef@petrochina.com.cn (X.Y.); liuyy93@petrochina.com.cn (Y.L.)
* Correspondence: lxz69@petrochina.com.cn (X.L.); zhanhongming17@mails.ucas.ac.cn (H.Z.)

Abstract: To address the limitations of traditional methods in modeling complex nonlinear relationships in horizontal in situ stress prediction for shale reservoirs, this study proposes an integrated framework that combines well logging interpretation with machine learning to accurately predict horizontal in situ stress in shale reservoirs. Based on the logging data from five wells in the Luzhou Block of the Sichuan Basin (16,000 samples), Recursive Feature Elimination (RF-RFE) was used to identify nine key factors, including Stoneley wave slowness and caliper, from 30 feature parameters. Bayesian optimization was employed to fine-tune the hyperparameters of the XGBoost model globally. Results indicate that the XGBoost model performs optimally in predicting maximum horizontal principal stress (SHmax) and minimum horizontal principal stress (SHmin). It achieves R^2 values of 0.978 and 0.959, respectively, on the test set. The error metrics (MAE, MSE, RMSE) of the XGBoost model are significantly lower than those of SVM and Random Forest, demonstrating its precise capture of the nonlinear relationships between logging parameters and in situ stress. This framework enhances the model's adaptability to complex geological conditions through multi-well data training and eliminating redundant features, providing a reliable tool for hydraulic fracturing design and wellbore stability assessment in shale gas development.

Keywords: horizontal in situ stress prediction; XGBoost; well logging interpretation; machine learning; complex nonlinearity; Bayesian optimization; shale reservoirs

check for updates

Academic Editor: Andrea Carpinteri

Received: 25 May 2025
Revised: 11 June 2025
Accepted: 16 June 2025
Published: 18 June 2025

Citation: Yu, W.; Li, X.; Guo, W.; Zhan, H.; Yang, X.; Liu, Y.; Pei, X.; He, W.; Wang, L.; Lin, Y. Prediction of Horizontal in Situ Stress in Shale Reservoirs Based on Machine Learning Models. *Appl. Sci.* **2025**, *15*, 6868. https://doi.org/10.3390/app15126868

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As global fossil fuel consumption keeps rising, unconventional natural gas has become crucial in exploration and development and will drive future oil and gas production growth [1–3]. The marine shale gas in Sichuan Basin, a key source of China's natural gas increase, is under wide attention for its development potential [4–8]. Stress field research is vital for shale gas development. horizontal stress distribution and magnitude affect wellbore stability and hydraulic fracturing design and are key for optimizing drilling and production strategies [9–12].

In the last few decades, accurate prediction of horizontal in situ stress in shale reservoirs has been a hot topic in geomechanics and reservoir engineering [13–16]. Conventional

well-logging interpretation has much experience in assessing underground geological properties. However, due to the high nonlinearity and complexity of geological structures, these methods cannot fully capture the factors affecting stress distribution [17–20]. In recent years, advanced computational methods have emerged. Machine-learning algorithms like XGBoost [21], SVM [22], and RF [23] excel in capturing complex nonlinear relationships and processing high-dimensional datasets. They have been successfully applied in hydrocarbon reservoir prediction and geomechanical parameter estimation.

However, existing research on horizontal in situ stress prediction for shale reservoirs has several limitations. Some studies rely solely on single-well data, which restricts the model's universality and generalization ability [24–26]. In feature selection, most works fail to adequately explore the deep-level correlations among different well-logging curves. They only focus on extracting morphological features of well-logging curves and finding nonlinear relationships, neglecting effective screening of key controlling factors [27,28]. Traditional model optimization methods often cannot achieve global optimality within limited data and computational resources, leading to instability in model performance across different datasets [29–31].

To address these gaps, this study introduces an innovative approach that integrates conventional well logging interpretation with advanced machine learning techniques for predicting horizontal in situ stress in shale reservoirs. Well logging interpretation provides a reliable dataset and theoretical guidance for the research, while XGBoost, known for its powerful performance, offers accurate and efficient data processing capabilities. The approach consists of three main steps: Initially, using the dataset from well logging interpretation, Pearson correlation coefficients and mutual information analysis are combined with the RF-RFE algorithm to identify the key factors influencing in situ stress. Subsequently, a prediction model is developed with these key factors as inputs and in situ stress as the target output, and Bayesian optimization is employed to adjust the hyperparameters, enhancing the model's global optimization capabilities. Finally, the performance of multiple machine learning models is compared to validate the proposed method's superiority. By addressing the shortcomings of prior studies, this research aims to provide a superior-performing model for in situ stress prediction in shale gas development and to establish a robust theoretical and practical foundation for future related studies.

The remainder of this paper is organized as follows: Section 2 outlines the experimental datasets and methods, covering the data source, characteristics, preprocessing, logging interpretation, XGBoost algorithm, RF-RFE feature selection, Bayesian optimization, and baseline models. Section 3 details the machine learning model, including data preprocessing steps and model evaluation metrics. Section 4 presents the results and discussion, comparing different models' prediction performances and exploring their field application implications. Finally, Section 5 summarizes the key findings and suggests future research directions.

2. Experimental Datasets and Methods

2.1. Experimental Datasets

The dataset utilized in this study comprises the logging curves of five wells from the 206H platform in the Luzhou Block, Sichuan Basin. Each well contains 32 logging curves, including depth (DEPTH), compensated acoustic wave (AC), caliper (CAL), compensated neutron (CNL), compensated density (DEN), natural gamma ray (GR), brittleness index (BRIT), transverse wave slowness (DTS), longitudinal wave slowness (DTC), Stoneley wave slowness (DTST), shear modulus (GMOD), bulk modulus (KMOD), true vertical depth (TVD), collapse pressure (BP), fracture pressure (PF), Poisson's ratio (POIS), pore pressure (PP), overburden pressure (SV), velocity ratio of transverse and longitudinal waves (VPVS),

Young's modulus (YMOD), potassium (K), thorium (TH), uranium (URAN), uranium-free gamma ray (KTH), photoelectric index (PE), deep lateral resistivity (RT), shallow lateral resistivity (RXO), continuous well temperature (TEMP), upper limit of safe mud density (PM1), lower limit of safe mud density (PM2), maximum horizontal principal stress (SHMAX), and minimum horizontal principal stress (SHMIN). In the training dataset, all 32 logging curves are utilized as input features. In the testing dataset, thirty logging curves, including compensated acoustic wave, caliper, natural gamma ray, and collapse pressure data, are used as input variables, with the maximum and minimum horizontal principal stresses as the target variables. The target variables SHMAX and SHMIN are derived from a calculation method based on dynamic elastic parameters. This method relies on formation parameters such as P-wave and S-wave transit times and density obtained from well logging data. Additionally, dynamic-to-static conversion coefficients acquired through laboratory tests are used to accurately determine the in situ stress values. Furthermore, the results are calibrated using on-site measurement data. The final SHMAX and SHMIN values are then utilized for predictions in machine learning models. The training and testing datasets jointly form the dataset for developing the horizontal principal stress prediction model for the Luzhou Block. Detailed information about the division of the training and testing datasets will be provided in the subsequent data pre-processing section.

In this study, well logging datasets from five wells are utilized. The depth range of the data is focused on the main shale gas-producing reservoir, spanning from 4000 to 6000 m. With 400 m of data extracted from each well at a sampling interval of 0.125 m, a total of 16,000 data samples covering a depth interval of 2000 m were obtained. Each well's data represents specific depth intervals, ensuring representativeness and enhancing dataset diversity. Using data from five wells instead of a single well aims to improve the model's generalization and adaptability. In geological exploration, logging data from different wells can vary due to factors like geological conditions and lithological properties. Training a model on data from a single well may lead to overfitting, limiting its applicability to other wells. By integrating data from five wells, this study reduces deviations caused by well-specific differences. This approach ensures the model's robust predictive performance across varying geological conditions, thereby enhancing its universality and reliability for practical applications.

2.2. Experimental Methodology

2.2.1. Logging Interpretation Method

Well logging interpretation involves analyzing the physical properties of subsurface formations to infer rock composition, porosity, permeability, and lithology. It provides fundamental data and theoretical support for geological analysis and in situ stress prediction [32–35]. Logging is conducted during or after drilling by lowering logging tools (sensors) into the well to measure and record various parameters of subsurface formations, such as resistivity, acoustic wave slowness, density, and gamma rays. Figure 1 uses data like acoustic wave slowness and density logging to calculate rock mechanical parameters such as elastic modulus and Poisson's ratio, which directly affect the magnitude and distribution of in situ stress. Resistivity logging helps estimate formation porosity and water saturation, which are closely related to the behavior of rock under in situ stress.

In the process of predicting in situ stress, well-logging interpretation supplies fundamental data to pinpoint the key factors affecting in situ stress. By examining various logging datasets, a comprehensive analysis of the physical properties of subsurface formations can be conducted, and characteristic parameters linked to in situ stress can be identified.

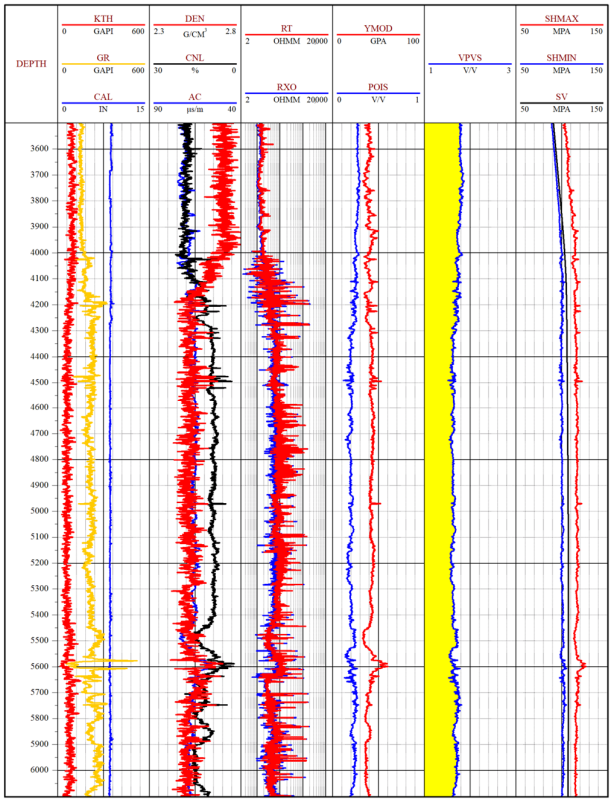


Figure 1. Logging Curve Chart of Rock Mechanics Parameters.

2.2.2. XGBoost

XGBoost is an efficient implementation of Gradient Boosting Decision Trees (GBDT) and is widely used for regression, classification, and ranking tasks [36,37]. Its high efficiency, flexibility, and excellent predictive performance make it a popular tool in machine learning, especially for complex problem-solving.

XGBoost aims to reduce model error by combining multiple decision trees. In each iteration, it adjusts the tree structure based on gradient information to minimize the loss function. This is achieved by optimizing an objective function, which is typically a weighted sum of a loss function and a regularization term, as shown in equation:

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

The loss function ($\ell(y_i, \hat{y}_i)$ measures the difference between the model's predicted values ($\hat{y}_i$) and the true values (y_i). Common loss functions include Mean Squared Error (MSE) and log loss. The regularization term ($\Omega(f_k)$) controls model complexity to prevent overfitting and is related to tree depth and the number of leaf nodes. K represents the number of trees.

In regression tasks, XGBoost commonly uses Mean Squared Error (MSE) as the loss function:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \tag{2}$$

In XGBoost, each iteration trains a new tree to minimize the current residual. To boost performance, XGBoost uses gradient boosting. Each new tree aims to reduce the previous model's error. In each iteration, XGBoost calculates the loss function's gradient based on current predictions (residuals) and adjusts model parameters using this gradient info. During optimization, XGBoost computes the first-order derivative (gradient) and the

second-order derivative of the objective function to adjust the tree structure. The gradient of the objective function for each sample is calculated as follows:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \tag{3}$$

The second-order derivative is calculated as follows:

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{4}$$

These gradients and second derivatives reflect the model's error at each point, guiding subsequent learning. In each iteration, XGBoost updates the model using the following formula:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{5}$$

In the formula, $f_t(x_i)$ represents the tree learned in the t-th iteration, and η denotes the learning rate, which controls each tree's contribution to the final model.

In XGBoost, each tree is built through a series of split nodes. The selection of split nodes is based on maximizing the reduction of the loss function after each split. Specifically, XGBoost uses a greedy algorithm to choose the optimal feature and split point at each node to divide the data. The gain from each split is calculated using the following formula:

$$Gain(t) = \frac{1}{2}\left(\frac{(\sum_L g_i)^2}{\sum_L h_i + \lambda} + \frac{(\sum_R g_i)^2}{\sum_R h_i + \lambda} - \frac{(\sum_t g_i)^2}{\sum_t h_i + \lambda}\right) - \gamma \tag{6}$$

In the formula, $\sum_L g_i$ and $\sum_R g_i$ represent the sum of gradients and second-order derivatives of the left and right child nodes, respectively. $\sum_t g_i$ represents the sum of gradients and second-order derivatives at the current node. λ and γ are regularization parameters that control the complexity of the tree. Using this gain formula, XGBoost selects the optimal split nodes to maximize the gain and constructs the tree structure.

Regularization is a key component of XGBoost, helping prevent overfitting, especially with high-dimensional data. The regularization term $\Omega(f_k)$ in XGBoost is typically defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

T is the number of leaf nodes of a tree, and γ controls the complexity of the number of leaf nodes per tree; w_j is the weight of the j-th leaf node, and λ is the L2 regularization term.The ranges of parameters λ and γ used in the model for this study are from 0.1 to 1.0. By introducing regularization, XGBoost can control the complexity of the tree, thereby enhancing the model's ability to generalize.

2.2.3. RF-RFE Algorithm

In feature selection, identifying which features most significantly impact horizontal in situ stress is crucial for developing a prediction model. In well logging curve interpretation, different feature values can be influenced by various potential factors, affecting the final interpretation. To select the key factors influencing horizontal in situ stress, this study combines Recursive Feature Elimination (RFE) with a Random Forest Regression (RF) model [38] to further screen and optimize features. This process removes redundant features and enhances the model's generalization ability. The Random Forest not only offers accurate predictions but also aids in understanding the relative importance of each feature in in situ stress prediction.

2.2.4. Bayesian Optimization

Bayesian optimization (BO) is an efficient global optimization method suitable for optimizing objective functions that are computationally expensive or difficult to analyze [39,40]. Unlike traditional optimization methods, BO constructs a surrogate model of the objective function and incrementally explores the input space to find the global optimum within a limited number of evaluations.

Bayesian optimization is widely used in hyperparameter tuning for machine learning. Hyperparameters are crucial for model performance, but their large and complex search spaces make efficient exploration difficult with conventional methods like grid or random search. Bayesian optimization uses a surrogate model to guide the search process, reducing the number of evaluations needed and improving efficiency. In the XGBoost, SVM and RF models, Bayesian optimization improves the model performance by tuning several hyperparameters and their value ranges, including: learning rate (0.01 to 0.05), number of trees (integer between 150 and 200), maximum depth of the tree (integer between 3 and 8), minimum loss reduction (0 to 5), sub-sampling ratio (0.5 to 0.8), column sampling ratio (0.5 to 0.8), and column sampling ratio (0.5 to 0.8). This process identifies the optimal model configuration.

2.2.5. Baseline Models

Random Forest (RF) is an ensemble learning method that predicts by constructing multiple decision trees and aggregating their results [41,42]. It uses bootstrapping to sample the original dataset with replacement, creating a distinct subset for each tree. During tree construction, nodes split on features randomly selected from a subset, enhancing tree independence and diversity. This method is robust, handles high-dimensional data and large datasets well, and is tolerant to missing and abnormal values. It also ranks feature importance, aiding in understanding their contribution to predictions. In this study, RF serves as a baseline model, offering an ensemble-based performance reference with rich parameters for subsequent tuning and comparison.

Support Vector Machine (SVM) is a supervised learning algorithm. Its core idea is to find an optimal hyperplane in feature space to maximize the separation between different classes [43]. For linearly separable data, SVM solves a convex quadratic programming problem to find this hyperplane. For linearly non-separable data, SVM uses kernel functions to map low-dimensional features to a high-dimensional space, making the data linearly separable. Common kernels include linear, polynomial, and radial basis function kernels, each suited to different data types and problems. SVM excels in handling small samples and high-dimensional data, has strong generalization ability, and effectively avoids overfitting. In this study, SVM is chosen as a baseline model due to its good performance in classification and regression tasks. Its decision boundaries, based on geometric and optimization theories, offer a predictive perspective different from Random Forest.

All baseline models are optimized using Bayesian optimization with 5-fold cross-validation conducted solely on the 80% training set. Performance evaluation and analysis of these baseline models help better understand the data and problem, providing a reference for comparing the improved model and assessing its performance gains.

2.2.6. Software Statement

In this study, the analysis was conducted using Python 3.8.10. The primary software libraries and packages utilized include pandas 1.3.5 for data manipulation, numpy 1.21.5 for numerical operations, scikit-learn 1.0.2 for machine learning models and utilities, XGBoost 1.5.2 for gradient boosting algorithms, matplotlib 3.5.1 for data visualization, and scikit-optimize 0.8.1 for Bayesian optimization.

3. Machine Learning Model

3.1. Data Preprocessing

Data preprocessing begins with data cleaning, which involves checking and converting inconsistent or invalid values in the raw data to ensure consistency and quality. Data are first converted into the appropriate type, with unconvertible parts marked as missing (NaN). This step identifies and addresses potential data format issues. Subsequently, features and labels are separated to ensure a clear and logical data structure for subsequent analysis and model training. For missing value handling, a tailored approach is adopted. Numerical features are filled using the mean value to maintain the overall data distribution, and categorical features are filled using the mode to preserve their representativeness. Outliers were detected using the interquartile range (IQR) method, with values below Q1 − 1.5 × IQR or above Q3 + 1.5 × IQR considered potential outliers. Instead of modifying or estimating outliers via methods like winsorization or imputation, we replaced them with a predefined value (9999) and removed rows containing this value or NaNs. This conservative approach aimed to preserve the integrity of non-outlier data and avoid potential biases from estimation. However, it might cause data loss, especially with many outliers. We acknowledge this trade-off and suggest future studies explore other outlier-handling strategies for a balance between data retention and quality. The resulting dataset has 13,304 entries. Eighty percent of the data (about 10,643 samples from the first four wells) forms the training set, and twenty percent (around 2661 samples from the fifth well) serves as the test set. This dataset is sufficient for training machine learning models and ensures data quality and reliability. By eliminating the impact of missing and abnormal values, the trained model achieves higher accuracy, stability, and generalization ability for more precise practical predictions.

To enhance model training and generalization, data normalization is essential. This study uses min–max normalization to map different feature curves to the [0, 1] interval. This eliminates the impact of varying feature scales on prediction accuracy. The formula is as follows:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{8}$$

In the formula, x_{min} represents the minimum value of the sample data, and x_{max} represents the maximum value of the sample data.

3.2. Model Evaluation Metrics

In the task of predicting in situ stress, metrics R^2, MAE, MSE, and RMSE were introduced to measure how well the models perform. R^2, which tells us how good the model fits the data, indicates the alignment between predicted and actual values. A model with R^2 closer to 1 has better predictive power. MAE, the average of absolute differences between predicted and true values, reflects the overall error size. It is less affected by outliers and directly shows how far off the predictions are in absolute terms. A smaller MAE means higher prediction accuracy. MSE, the average of squared differences between predicted and true values, assesses both the overall error and its dispersion. RMSE, being the square root of MSE, shares the same unit as the target variable, making it more intuitive. These metrics together allow for a comprehensive evaluation of the SVM, RF, and XGBoost models. When comparing models, R^2 is the primary measure for assessing how well the model fits the data, while MAE, MSE, and RMSE provide additional perspectives on the prediction errors. The best-performing model, as determined by these metrics, is chosen for practical applications.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - y_i'\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^{'}\right| \quad (10)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2 \quad (11)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2} \quad (12)$$

3.3. Analysis of Dominating Factors

Identifying the features with the greatest impact on horizontal in situ stress is crucial for model development. During actual well logging curve interpretation, various features may be influenced by numerous potential factors. This study combines Pearson correlation analysis with mutual information evaluation to assess feature correlations and their relevance to target variables. Recursive Feature Elimination and Random Forest regression are then used to screen and optimize features, eliminating redundancy and enhancing model generalization.

First, the Pearson correlation coefficient matrix is calculated to measure linear relationships among features. Features with correlations exceeding the 0.6 threshold are flagged as potential redundant features. As shown in Figure 2a, this matrix provides a clear visualization of feature dependencies and highlights clusters of highly correlated features.

For features with high Pearson correlation, the mutual information with the two target variables (maximum and minimum horizontal principal stress) is calculated. Mutual information in Table 1 quantifies the nonlinear dependence between features and targets, helping identify features with unique predictive contributions.

Table 1. Mutual Information Values Between Feature Values and Target Variables.

Features	Mutual Information Value	Features	Mutual Information Value	Features	Mutual Information Value
DEPTH	1.816	TVD	1.234	SV	1.216
PP	1.195	BDYN	1.189	AC	0.987
YMOD	0.879	TEMP	0.862	DTC	0.862
DTS	0.721	PF	0.687	GDYN	0.623
DTST	0.617	CAL	0.584	VPVS	0.563
PM1	0.561	POIS	0.558	BRIT	0.475
CNL	0.428	GR	0.337	BP	0.283
RT	0.255	TH	0.248	URAN	0.210
PM2	0.209	DEN	0.208	RXO	0.200
KTH	0.188	PE	0.172	K	0.114

A mutual information value between 0 and 0.3 indicates a weak correlation between the feature and the target variable. A value between 0.3 and 0.6 suggests a moderate correlation. A value between 0.6 and 1 implies a strong correlation. And a value greater than 1 shows an over-correlation between the feature and the target variable. Higher mutual information values indicate greater feature contributions to the target variables. However, features with over-correlation (mutual information > 1) may indicate multicollinearity, carrying similar information. Models relying heavily on them may overfit and underperform on new data. Our feature selection priority based on mutual information is strong (0.6–1) > moderate (0.3–0.6) > weak (0–0.3) > over-correlation (>1). Thus, the mutual information threshold is set at 1. Features exceeding this threshold are considered redundant and are prioritized for removal. For pairs of highly correlated features (with Pearson correlations above the threshold), if both have mutual information below 1, the

feature with higher mutual information is retained. If one feature has mutual information above 1, that feature is removed. This process reduces redundancy while retaining predictive features. After iterative screening, the feature set is refined, and the Pearson correlation matrix is recalculated and analyzed to verify the elimination of redundancy. As shown in Figure 2b, the correlations in the refined matrix are significantly reduced, demonstrating the effectiveness of the feature selection method.

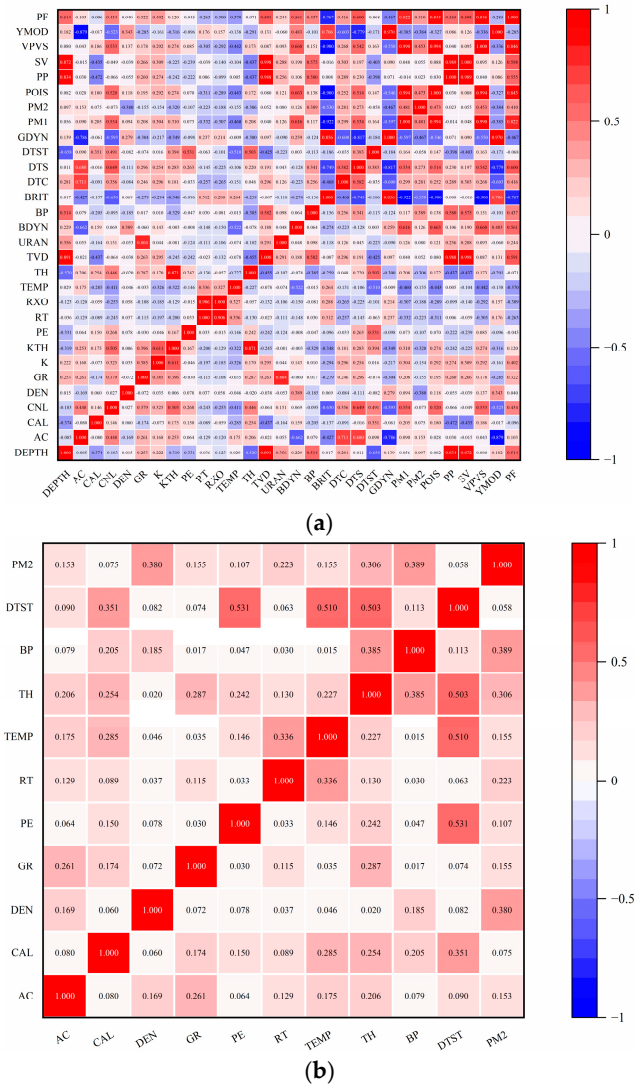


Figure 2. Feature Correlation Coefficient Heat Matrix. (**a**) The correlation coefficient matrix of the initial 30 features. (**b**) The matrix after screening, displaying correlations among the remaining 11 features.

Recursive Feature Elimination with Cross-Validation (RFECV) and Random Forest regression were used for feature screening and optimization. To prevent overfitting, the Random Forest's structure was constrained by limiting the maximum tree depth, increasing the minimum samples per leaf node, and reducing the number of trees. The model achieved a high R^2 on the training set and good prediction performance on the test set. Feature importances from the model were visualized in Figure 3. DTST had the highest relative importance (0.3524), followed by CAL (0.1884) and TTEMP (0.1525), indicating their significant impact on predicting the target variables. RFECV with 10-fold cross-validation was then applied to further optimize the feature set. RFECV removes features with low model contributions and evaluates different feature combinations via cross-validation to determine the optimal number of features. The selected feature set balances model accuracy

and simplicity. As shown in Figure 4, the RFECV cross-validation results indicate that the model's average R^2 score peaks when the number of features approaches 9, indicating the best prediction performance with this feature combination. The feature selection results, including rankings, selection markers, and cross-validation scores, are stored in the output directory. The RFECV process is also visualized, showing the model's R^2 value trends with different feature numbers, supporting the rationality of the Recursive Feature Elimination.

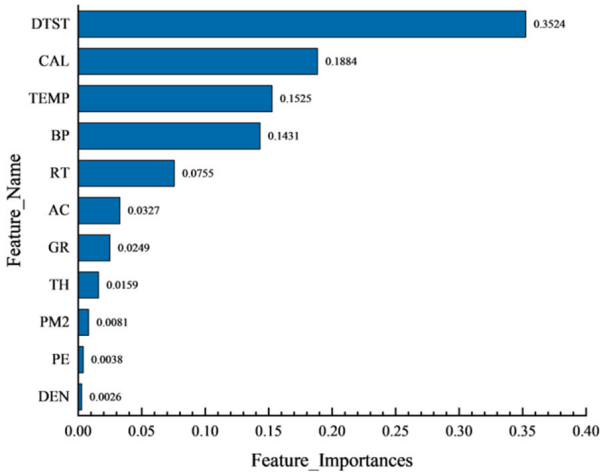


Figure 3. Distribution of Feature Importances After Screening.

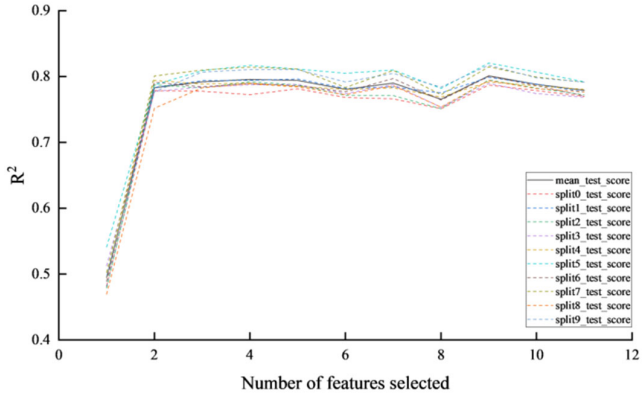


Figure 4. Changes in R^2 Test Scores with Feature Selection Using Recursive Feature Elimination and Cross-Validation.

4. Results and Discussion

4.1. Comparison of Different Models' Prediction Results

To predict the magnitude of horizontal principal stress in the Luzhou area, this study selected three machine learning models: XGBoost, SVM, and RF. These models were trained using the nine key factors selected by the RF-RFE algorithm as inputs, with the maximum and minimum horizontal principal stresses as target values. During training, Bayesian optimization was used to determine the optimal hyperparameters. The three models were then compared based on simulation results and evaluation metrics to identify the best one.

The nine selected factors were used to train and optimize the models. As shown in Figure 5, the XGBoost and SVM models predicted horizontal principal stress values closer to the actual values, while the RF model showed larger deviations in some data points. However, it was difficult to determine which of the XGBoost or SVM models had predictions closer to the true values based on the average.

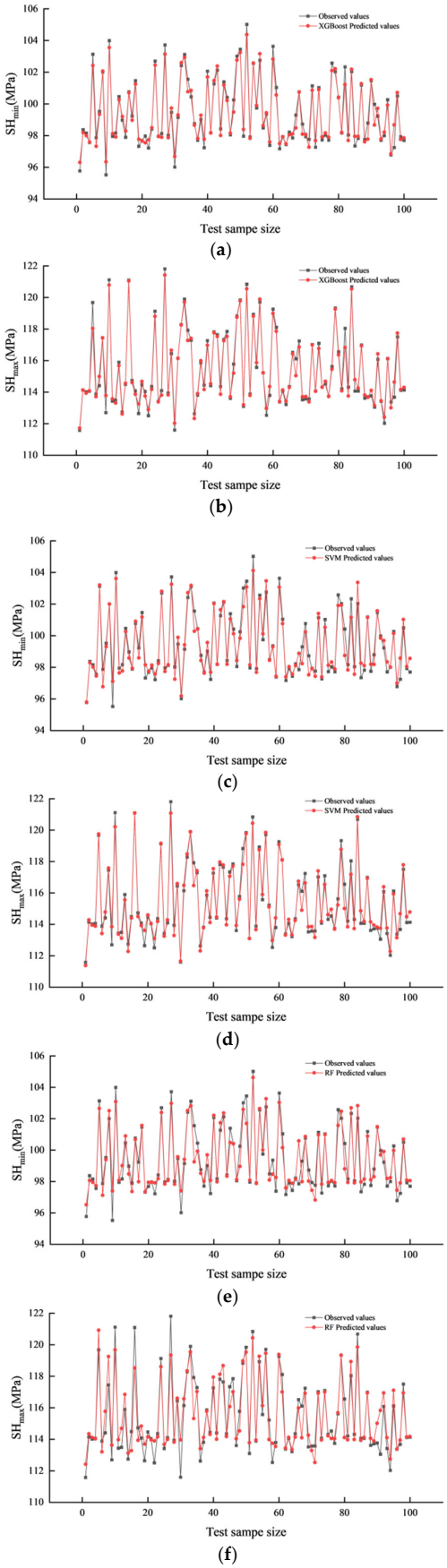


Figure 5. Comparison of Predicted and Observed Values for Different Models. (**a**,**b**) The predicted vs. observed values of SHmin and SHmax for the XGBoost model. (**c**,**d**) The values for the SVM model. (**e**,**f**) The results for the RF model.

To better observe the actual prediction performance of the models, scatter plots of actual versus predicted values along with three reference lines, were created to compare and evaluate the performance of the XGBoost, SVM, and RF models in predicting the SHmax and SHmin, as shown in Figure 6. The red dashed line indicates where the predicted values match the actual values, while the green and blue dashed lines represent the ±1% error bounds. The scatter points of the XGBoost model are tightly clustered around the red dashed line and within the ±1% error bounds, indicating that its predictions are highly accurate and stable with minimal deviation from the actual values. The SVM model's scatter points also show a certain degree of regularity, with most points lying within the ±1% error bounds despite some dispersion, suggesting strong prediction capability and stability. In contrast, the RF model's scatter points exhibit a higher degree of dispersion and are less constrained by the ±1% error bounds, indicating relatively weaker prediction accuracy and stability with potential prediction biases.

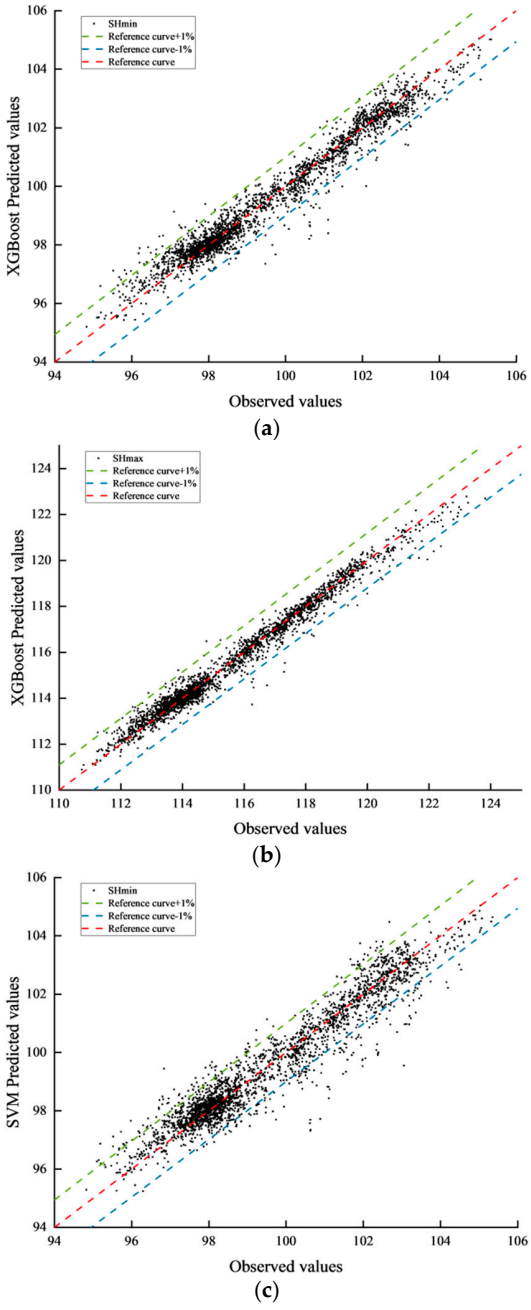


Figure 6. *Cont.*

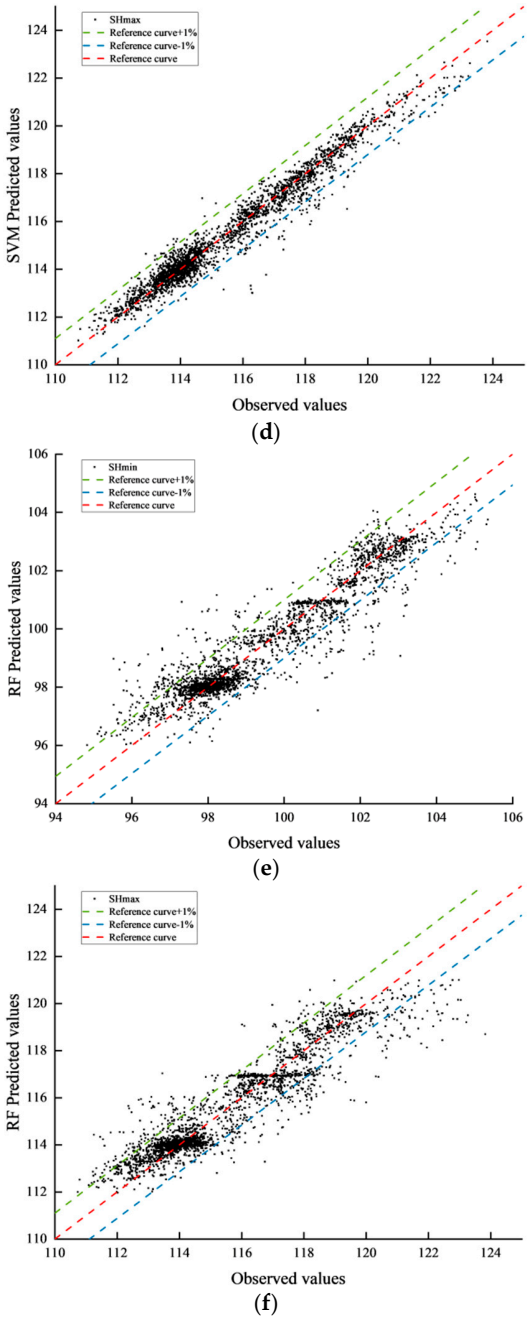


Figure 6. Comparison of Prediction Accuracy and Stability of Different Models. (**a**,**b**) Scatter plots for XGBoost's SHmin and SHmax predictions within ±1% error. (**c**,**d**) The same for SVM. (**e**,**f**) are for RF.

4.2. Comparison of Different Models Evaluation Metrics

The evaluation of the XGBoost, SVM, and RF models for predicting SHmax and SHmin reveals significant differences in their accuracy and robustness. As shown in Table 2, XGBoost outperforms the other models in SHmax prediction with MAE, MSE, and RMSE values of 0.250, 0.125, and 0.353, respectively. These values are markedly lower than those of SVM (MAE = 0.345, MSE = 0.242, RMSE = 0.492) and RF (MAE = 0.621, MSE = 0.763, RMSE = 0.874). XGBoost also achieves an R^2 of 0.978, indicating it can explain 97.8% of the variance in SHmax data, outperforming SVM (R^2 = 0.958) and RF (R^2 = 0.866). In SHmin prediction, XGBoost continues to lead with the lowest MAE, MSE, and RMSE values (0.304,

0.181, and 0.426) and the highest R^2 value (0.959). While RF's R^2 for SHmin (0.885) is slightly better than for SHmax, it still trails behind XGBoost and SVM.

Figure 7 further validates these conclusions through a comprehensive analysis of multiple metrics via radar charts. In the chart, a model demonstrates superior comprehensive performance if its R^2 dimension is closer to the outer edge and its MAE, MSE, and RMSE dimensions are closer to the center. XGBoost demonstrates near-maximum R^2 values and near-minimum error values, highlighting its superior prediction accuracy and goodness of fit. SVM, while outperforming RF, shows higher error metrics and lower R^2 values compared to XGBoost, indicating less favorable predictive performance. This suggests that SVM, though capable in nonlinear regression tasks, is not as effective as XGBoost in this specific application. Conversely, RF exhibits the largest radar chart coverage, with higher error metric dispersion and the smallest R^2 value. This indicates RF's predictions are more sensitive to outliers and have weaker generalization performance. Overall, XGBoost, with its integrated learning mechanism and hyperparameter optimization strategy, shows higher reliability and engineering applicability in the nonlinear mapping of complex geological features and target variables, and provides an optimal solution for horizontal in situ stress prediction.

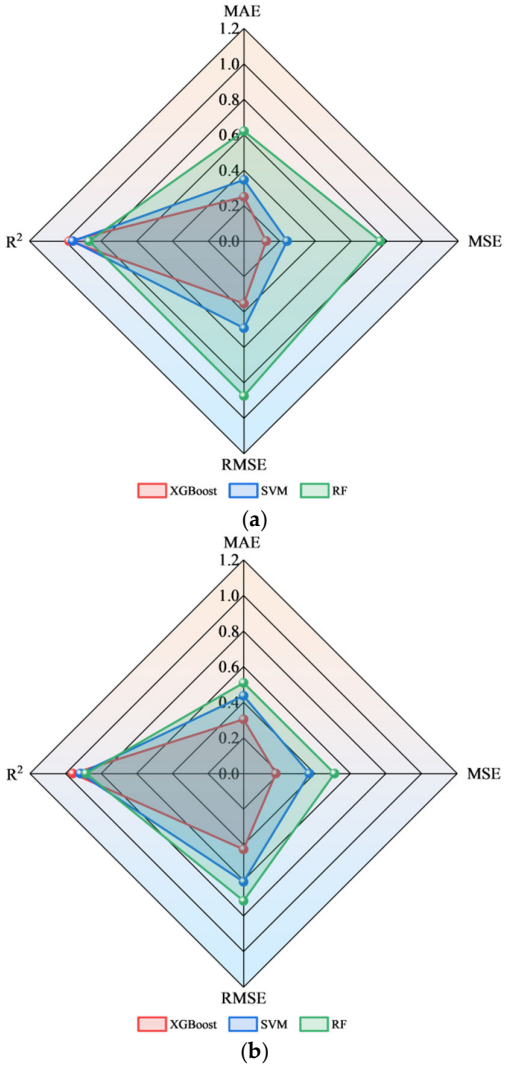


Figure 7. Radar Chart of Performance Metrics for XGBoost, SVM, and RF Models. (**a**) Evaluation metrics for SHmax prediction. (**b**) Evaluation metrics for SHmin prediction.

Table 2. Evaluation Metrics (R^2, MAE, MSE, RMSE) of XGBoost, SVM, and RF Models for SHmax and SHmin Prediction on 20% Unseen Test Set.

Target Variables	Parameters	XGBoost	SVM	RF
SHmax	MAE	0.250	0.345	0.621
	MSE	0.125	0.242	0.763
	RMSE	0.353	0.492	0.874
	R^2	0.978	0.958	0.866
SHmin	MAE	0.304	0.437	0.509
	MSE	0.181	0.370	0.511
	RMSE	0.426	0.608	0.715
	R^2	0.959	0.917	0.885

4.3. Field Application

In this study, across the entire test set, the XGBoost model predicted a maximum horizontal principal stress of 110.89~122.78 MPa and a minimum horizontal principal stress of 95.21~105.05 MPa, with a horizontal stress difference of 12.11~20.53 MPa. The actual stress log of a well on site shows that in the 4230.0~5907.5 m section, the maximum horizontal principal stress ranges from 112.3 to 126.5 MPa, the minimum horizontal principal stress ranges from 96.6 to 106.9 MPa, and the differential stress between the maximum and minimum horizontal principal stresses is 16.3 MPa. The predicted ranges of maximum and minimum horizontal principal stresses exhibit errors within 5%, aligning with engineering tolerance standards. These data not only quantify the stress distribution in the target area but also offer key references for on-site operations. They help optimize hydraulic fracturing designs, predict fracture propagation, and assist in geological modeling, thereby enhancing shale gas development efficiency.

Hydraulic fracturing, which depends heavily on stress distribution, focuses on fracture shape, size, and network uniformity in its design. The predicted results indicate that the horizontal stress difference in the Luzhou area is moderate, and fracture propagation is mainly controlled by the maximum horizontal principal stress direction. A lower stress difference (12.11 MPa) may lead to a complex fracture network, forming multi-branch fractures and helping to expand the reservoir modification range. In contrast, a higher stress difference (20.53 MPa) makes fractures tend toward a single main-fracture pattern, with longer fracture lengths but lower fracture complexity. Therefore, in areas with low stress differences, high-viscosity cross-linked fracturing fluids should be used to enhance fracture network complexity and improve reservoir sweep efficiency. In areas with high stress differences, slick water fracturing is more suitable to ensure effective fracture propagation along the maximum horizontal principal stress direction. Additionally, perforation cluster spacing should be optimized based on the stress field, with spacing reduced in areas of high stress gradients to improve fracture uniformity and prevent single fractures from dominating the entire fracturing stage.

Fracture propagation is controlled by the stress difference. In areas with small horizontal stress differences, complex fracture networks are more likely to form, while large stress differences favor the generation of stable main fractures. Based on the predicted data, the fracture propagation pattern in the Luzhou area can be further optimized. In areas with low stress differences, the complex fracture network helps improve reservoir sweep efficiency, but fracture widths may be narrow, limiting proppant transportation. Thus, the fracturing fluid system needs to be optimized to enhance sand-carrying capacity. In areas with high stress differences, fractures tend to be single, requiring control of fracturing fluid viscosity to prevent excessive fracture width and insufficient proppant filling. The stress field distribution also determines fracture height: in areas with large horizontal

stress differences, fractures tend to propagate horizontally, while in areas with small stress differences, fractures may exhibit some vertical propagation. Therefore, fracture height control should be combined with the stress field to optimize fracturing designs.

Geomechanical modeling is crucial for optimizing reservoir development. Based on stress prediction results, geomechanical models can be refined to enhance wellbore stability and completion parameter optimization. The minimum horizontal principal stress distribution in the Luzhou area indicates that wellbores in some regions, especially in low-stress areas (SHmin = 95.21 MPa), may be at higher risk of collapse. In low-stress areas, high-density mud (>1.4 g/cm^3) should be used to enhance wellbore support, and high-collapse-strength casing materials should be selected in completion designs. In high-stress areas, casing deformation needs attention, and appropriate completion parameters should be chosen to improve long-term wellbore stability. Moreover, the predicted results can be used for stress inversion, combining field-measured data to adjust stress distribution models and enhance their reliability. This provides more precise guidance for subsequent well location optimization and drilling and completion operations.

5. Conclusions

1. RFECV identified 9 key features from 30 initial well-logging curves, eliminating 21 weakly related ones to reduce redundancy. This enhances the model's interpretability and generalization in heterogeneous geological conditions. Bayesian optimization effectively adjusts hyperparameters like tree depth and learning rate, improving model convergence for a globally optimal configuration. Compared to grid and random search, this method reduces computational cost and ensures robustness across datasets.
2. The XGBoost model outperforms baseline models (SVM and RF) in SHmax and SHmin prediction. It achieved the highest R^2 (0.978 for SHmax, 0.959 for SHmin) and lowest error indicators (MAE, MSE, RMSE). This highlights its ability to capture the nonlinear relationship between logging parameters and stress distribution. However, the model's applicability to other shale formations has not been adequately researched, which may introduce certain limitations. Future studies can validate the model's performance across different shale formations to enhance its generalizability and reliability in varied geological settings.
3. The predicted SHmax and SHmin data reveal the reservoir's stress distribution, aiding in optimizing fracture propagation during hydraulic fracturing. In areas with low stress differences, adjusting fracturing parameters can enhance fracture uniformity. In areas with high stress differences, optimizing proppant carrying capacity ensures fracture conductivity.
4. The prediction results are useful for geomechanical modeling and wellbore stability design. In low-stress areas, high-density mud prevents wellbore collapse; in high—stress areas, optimizing casing strength improves well integrity. Field data like microseismic monitoring and stress logging can iteratively refine the prediction model, enhancing its applicability throughout shale gas development.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to some data confidentiality restrictions.

Conflicts of Interest: Authors Wenxuan Yu, Xizhe Li, Wei Guo, Hongming Zhan, Xiangyang Pei, Weikang He, Longyi Wang, Yaoqiang Lin were employed by the company Research Institute of Petroleum Exploration and Development, PetroChina. Authors Xuefeng Yang, Yongyang Liu, were employed by the company Shale Gas Research Institute of Southwest Oil and Gas Field Branch. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Li, G.X.; Lei, Z.D.; Dong, W.H.; Wang, H.Y.; Zheng, X.F. Progress, Challenges, and Prospects of Unconventional Oil and Gas Development in China. *China Pet. Explor.* **2022**, *27*, 1–11.
2. Li, L.G. Review and Prospects of China's Natural Gas Industry Development. *Nat. Gas Ind.* **2021**, *41*, 1–11.
3. Jia, A.L.; He, D.B.; Wei, Y.S.; Li, Y. Forecast of China's Natural Gas Development Trend in the Next 15 Years. *Nat. Gas Geosci.* **2021**, *32*, 17–27.
4. Ma, X.H. Theory and Practice of "Limit Recovery" Development of Unconventional Natural Gas. *Pet. Explor. Dev.* **2021**, *48*, 326–336. [CrossRef]
5. Hu, W.R.; Li, H.T.; Wang, J.R.; Pan, Z.Y.; Dang, L.R.; Jing, X.S.; Bao, J.W.; Huang, M.; Yu, G.; Zhu, H.; et al. Strategy, Implementation Path and Safeguard Measures of "Large Gas Daqing" in Southwest China. *Nat. Gas Ind.* **2023**, *43*, 146–151.
6. Ma, X.H.; Zhang, X.W.; Xiong, W.; Liu, Y.Y.; Gao, J.L.; Yu, R.Z.; Sun, Y.P.; Wu, J.; Kang, L.X.; Zhao, S.P. Development Prospects and Challenges of Shale Gas in China. *Bull. Pet. Sci. Technol.* **2023**, *8*, 491–501.
7. Ni, H.K.; Dang, W.; Zhang, K.; Su, H.K.; Ding, J.H.; Li, D.H.; Liu, X.W.; Li, P.; Li, P.; Yang, S.Y.; et al. 20 Years of Shale Gas Research and Development in China: Review and Prospect. *Nat. Gas Ind.* **2024**, *44*, 20–52.
8. Zou, C.N.; Dong, D.Z.; Xiong, W.; Fu, G.Y.; Zhao, Q.; Liu, W.; Kong, W.L.; Zhang, Q.; Cai, G.Y.; Wang, Y.M.; et al. Progress, Challenges and Countermeasures of Shale Gas Exploration in New Zones, New Strata and New Types in China. *Oil Gas Geol.* **2024**, *45*, 309–326.
9. Li, M.; Liu, Y.L.; Feng, D.J.; Shen, B.J.; Du, W.; Wang, P.W. Potential of Marine Shale Gas Resources and Future Exploration Direction in China. *Exp. Pet. Geol.* **2023**, *45*, 1097–1108.
10. Yang, Z.Z.; Yuan, J.F.; Zhang, J.Q.; Li, X.G.; Zhu, J.Y.; He, J.K. Research Progress and Understanding of Fracture Networks in horizontal Wells of Marine Shale in the Sichuan Basin. *Pet. Reserv. Eval. Dev.* **2024**, *14*, 600–609.
11. Xusheng, G.U.O.; Ruyue, W.A.N.G.; Baojian, S.H.E.N.; Guanping, W.A.N.G.; Chengxiang, W.A.N.; Qianru, W.A.N.G. Geological Characteristics, Resource Potential and Development Direction of Shale Gas in China. *Pet. Explor. Dev.* **2025**, *52*, 15–28.
12. Tang, H.Y.; Luo, S.G.; Liang, H.P.; Zeng, B.; Zhang, L.H.; Zhao, Y.L.; Song, Y. Numerical Simulation of Integrated Fracturing-Production for Shale Gas Wells Considering Gas-Water Two-Phase Flow. *Pet. Explor. Dev.* **2024**, *51*, 597–607. [CrossRef]
13. Liu, Y.Y.; Ju, W.; Xiong, W.; Guo, W.; Ning, W.K.; Yu, G.D.; Liang, X.B.; Li, Y.K. Present-Day In-Situ Stress Characteristics and Shale Gas Development in the Luzhou Block, Southern Sichuan. *Sci. Technol. Eng.* **2024**, *24*, 3200–3206.
14. Han, L.L.; Li, X.Z.; Liu, Z.Y.; Duan, G.F.; Wan, Y.J.; Guo, X.L.; Guo, W.; Cui, Y. Main Controlling Factors and Countermeasures for Casing Deformation in Deep Shale Gas Wells in Southern Sichuan. *Pet. Explor. Dev.* **2023**, *50*, 853–861. [CrossRef]
15. Zhu, H.Y.; Song, Y.J.; Tang, X.H. Progress in Research on 4D In-Situ Stress Evolution in Shale Gas Reservoirs and Complex Fracture Propagation in Infilling Wells. *Bull. Pet. Sci. Technol.* **2021**, *6*, 396–416.
16. Shi, X.W.; Wang, C.; Zhang, D.J.; Du, B.Y.; Gao, J.H.; Dong, X.H.; Wu, T.; Zhang, J.X. Seismic Prediction Technology for In-Situ Stress in Deep Shale Gas Reservoirs of the Wufeng–Longmaxi Formation, Northern Luzhou Area, Sichuan Basin. *Nat. Gas Geosci.* **2024**, *35*, 2040–2052.
17. Du, B.Y.; Gao, J.H.; Zhang, G.Z.; Dong, X.H.; Guo, W.; Zhang, J.D. Seismic Prediction Method and Application of In-Situ Stress in Shale Reservoir Based on Fracture Density Inversion. *Oil Geophys. Prospect.* **2024**, *59*, 279–289.
18. Chen, X.H. Evaluation and Engineering Application of Present-Day In-Situ Stress Field of the Longmaxi Shale Formation in the Dingshan–Dongxi Area. Master's Thesis, Chengdu University of Technology, Chengdu, China, 2023.
19. He, C. Study on Prediction of In-Situ Stress and Compressibility of Shale Reservoir Based on VTI Medium. Master's Thesis, Chengdu University of Technology, Chengdu, China, 2023.
20. Cheng, D.J.; Sun, B.D.; Cheng, Z.G.; Wan, J.B.; Wang, H.; Zhang, Y.H. Current Status and Prospect of In-Situ Stress Evaluation Based on Logging Data. *Well Logging Technol.* **2014**, *38*, 379–383.

21. Yan, J.F.; Li, S.L.; Wei, Z.D.; Wu, Z.B.; Chen, J.Y. Lithofacies Prediction Method for Shale Based on XGBoost Algorithm. *J. Palaeogeogr.* **2025**, *27*, 763–776.
22. Ni, W.J.; Li, Q.; Guo, W.H.; Feng, T.; Li, X.M.; Zhou, T.T. Prediction of Shear Wave Velocity in Shale Reservoirs Based on Support Vector Machine. *J. Xi'an Shiyou Univ. (Nat. Sci. Ed.)* **2017**, *32*, 46–49+54.
23. Duan, Y.X.; Wang, Y.F.; Sun, Q.F. Application of Selective Ensemble Learning Model in Lithology-Porosity Prediction. *Sci. Technol. Eng.* **2020**, *20*, 1001–1008.
24. Liu, T.; Tian, R.F.; Zhang, W. Prediction of Logging Shear Wave Velocity Based on Deep Neural Network Optimized by Genetic Algorithm. *Geophys. Geochem. Explor. Comput. Technol.* **2023**, *45*, 289–298.
25. Luo, F.Q.; Liu, J.T.; Chen, X.P.; Li, S.A.; Yao, X.Z.; Chen, D. Intelligent Prediction of Formation Pore Pressure in the No. 5 Fault Zone of Shunbei Oilfield Based on BP and LSTM Neural Network. *Oil Drill. Prod. Technol.* **2022**, *44*, 506–514.
26. Ma, T.S.; Xiang, G.F.; Shi, Y.F.; Gui, J.C.; Zhang, D.Y. Horizontal Stress Prediction Method Based on Bidirectional Long Short-Term Memory Neural Network. *Bull. Pet. Sci. Technol.* **2022**, *7*, 487–504.
27. Ma, T.; Xiang, G.; Shi, Y.; Lui, Y. Horizontal In-Situ Stresses Prediction Using a CNN-BiLSTM-Attention Hybrid Neural Network. *Geomech. Geophys. Geo-Energy Geo-Resour.* **2022**, *8*, 152. [CrossRef]
28. Lin, H.; Kang, W.; Oh, J.; Canbulat, I. Estimation of In-Situ Maximum Horizontal Principal Stress Magnitudes from Borehole Breakout Data Using Machine Learning. *Int. J. Rock Mech. Min. Sci.* **2020**, *126*, 104199. [CrossRef]
29. Zhao, H.; Yin, S. Geomechanical Parameters Identification by Particle Swarm Optimization and Support Vector Machine. *Appl. Math. Model.* **2009**, *33*, 3997–4012. [CrossRef]
30. Ibrahim, A.F.; Gowida, A.; Ali, A.; Elkatatny, S. Real-Time Prediction of In-Situ Stresses While Drilling Using Surface Drilling Parameters from Gas Reservoir. *J. Nat. Gas Sci. Eng.* **2022**, *97*, 104368. [CrossRef]
31. Chang, Z.; Catani, F.; Huang, F.; Liu, G.; Meena, S.R.; Huang, J.; Zhou, C. Landslide Susceptibility Prediction Using Slope Unit-Based Machine Learning Models Considering the Heterogeneity of Conditioning Factors. *J. Rock Mech. Geotech. Eng.* **2023**, *15*, 1127–1143. [CrossRef]
32. Han, X.; Yi, X.; Zhou, C.; Che, X.; Hou, L.; Huang, X.; Wang, Y.; Zeng, J. Study on Rock Mechanics Parameters and In-Situ Stress Profile Construction and Correction Method Based on Well Log Interpretation. *Chem. Technol. Fuels Oils* **2021**, *57*, 518–528. [CrossRef]
33. Zhang, W.; Wang, J.; Li, K.; Liu, H.; Kang, Y.; Wu, Y.; Lv, W. Unilateral Alignment: An Interpretable Machine Learning Method for Geophysical Logs Calibration. *Artif. Intell. Geosci.* **2021**, *2*, 192–201. [CrossRef]
34. Saeed, A.; Hamidzadeh, R.M.; Ahsan, L. New Interpretation Approach of Well Logging Data for Evaluation of Kern Aquifer in South California. *J. Appl. Geophys.* **2023**, *215*, 105138.
35. Chen, H. Study on Interpretation Method of Remaining Oil Saturation Based on PSSL Logging. *Acad. J. Environ. Earth Sci.* **2023**, *5*, 55–62.
36. 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, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
37. Hu, W.; Liang, J.; Jin, Y.; Wu, F.; Wang, X.; Chen, E. Online Evaluation Method for Low Frequency Oscillation Stability in a Power System Based on Improved XGBoost. *Energies* **2018**, *11*, 3238. [CrossRef]
38. Díaz-Uriarte, R.; Alvarez de Andrés, S. Using Recursive Feature Elimination in Random Forest to Account for Correlated Variables in High Dimensional Data. *BMC Genet.* **2018**, *19*, 65.
39. Lima, F.C.; Lobo, G.F.; Pelikan, M.; Goldberg, D.E. Model Accuracy in the Bayesian Optimization Algorithm. *Soft Comput.* **2011**, *15*, 1351–1371. [CrossRef]
40. Monson, K.C.; Seppi, D.K. Bayesian Optimization Models for Particle Swarms. In Proceedings of the 7th Annual Conference on GENETIC and Evolutionary Computation, Washington, DC, USA, 25 June 2005; pp. 193–200.
41. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
42. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef]
43. Singh, U.B.; Nanhay, S. Analysis of Recent Advancements in Support Vector Machine. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7270.

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