

Article

Lithofacies Identification by an Intelligent Fusion Algorithm for Production Numerical Simulation: A Case Study on Deep Shale Gas Reservoirs in Southern Sichuan Basin, China

Yi Liu ^{1,2}, Jin Wu ^{3,4,*}, Boning Zhang ^{1,2}, Chengyong Li ⁵, Feng Deng ⁵, Bingyi Chen ^{1,2}, Chen Yang ^{1,2}, Jing Yang ^{1,2} and Kai Tong ^{1,2}

¹ Chengdu Northern Petroleum Exploration and Development Technology Co., Ltd., Chengdu 610051, China; liuyi@zhenhuaoil.com (Y.L.); zhangboning@zhenhuaoil.com (B.Z.); chenbingyi@zhenhuaoil.com (B.C.); yangchen@zhenhuaoil.com (C.Y.); yangjing@zhenhuaoil.com (J.Y.); tongkai@zhenhuaoil.com (K.T.)

² Zhenhua Oil Co., Ltd., Beijing 100031, China

³ PetroChina Research Institute of Petroleum Exploration and Development, Beijing 100083, China

⁴ National Energy Shale Gas Research and Development (Experiment) Center, Beijing 100083, China

⁵ Energy College, Chengdu University of Technology, Chengdu 610059, China; lichengyong1981@126.com (C.L.); qigeyou@163.com (F.D.)

* Correspondence: wujinouc@petrochina.com.cn

Abstract

Lithofacies, as an integrated representation of key reservoir attributes including mineral composition and organic matter enrichment, provides crucial geological and engineering guidance for identifying “dual sweet spots” and designing fracturing strategies in deep shale gas reservoirs. However, reliable lithofacies characterization remains particularly challenging owing to significant reservoir heterogeneity, scarce core data, and imbalanced facies distribution. Conventional manual log interpretation tends to be cost prohibitive and inaccurate, while existing intelligent algorithms suffer from inadequate robustness and suboptimal efficiency, failing to meet demands for both precision and practicality in such complex reservoirs. To address these limitations, this study developed a super-integrated lithofacies identification model termed SRLCL, leveraging well-logging data and lithofacies classifications. The proposed framework synergistically combines multiple modeling advantages while maintaining a balance between data characteristics and optimization effectiveness. Specifically, SRLCL incorporates three key components: Newton-Weighted Oversampling (NWO) to mitigate data scarcity and class imbalance, the Polar Light Optimizer (PLO) to accelerate convergence and enhance optimization performance, and a Stacking ensemble architecture that integrates five heterogeneous algorithms—Support Vector Machine (SVM), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM)—to overcome the representational limitations of single-model or homogeneous ensemble approaches. Experimental results indicated that the NWO-PLO-SRLCL model achieved an overall accuracy of 93% in lithofacies identification, exceeding conventional methods by more than 6% while demonstrating remarkable generalization capability and stability. Furthermore, production simulations of fractured horizontal wells based on the lithofacies-controlled geological model showed only a 6.18% deviation from actual cumulative gas production, underscoring how accurate lithofacies identification facilitates development strategy optimization and provides a reliable foundation for efficient deep shale gas development.



Academic Editors: Dicho Stratiev and Qiang Yu

Received: 26 October 2025

Revised: 28 November 2025

Accepted: 2 December 2025

Published: 14 December 2025

Citation: Liu, Y.; Wu, J.; Zhang, B.; Li, C.; Deng, F.; Chen, B.; Yang, C.; Yang, J.; Tong, K. Lithofacies Identification by an Intelligent Fusion Algorithm for Production Numerical Simulation: A Case Study on Deep Shale Gas Reservoirs in Southern Sichuan Basin, China. *Processes* **2025**, *13*, 4040. <https://doi.org/10.3390/pr13124040>

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/>).

Keywords: shale gas; lithofacies; polar lights optimization; oversampling enhancement; numerical simulation

1. Introduction

Lithofacies, an intrinsic property of shale, refers to the characteristics of a rock or rock assemblage formed under specific sedimentary environments [1]. It records the reservoir's depositional environment and developmental history [2]. By encapsulating shale's mineralogy and biotic signatures, lithofacies critically controls reservoir properties [3]. Different lithofacies exhibit significant heterogeneity in spatial distribution, physical properties, and pore structure characteristics, consequently demonstrating varying hydrocarbon generation potential and accumulation capacity [4]. As the foundation of exploration and stimulation strategies, lithofacies studies enable reservoir characterization, guiding identification of high-quality zones with optimal hydrocarbon enrichment and fracability [5,6].

Traditional qualitative methods for lithofacies identification, such as manual description and well-log interpretation, yield results whose reliability is heavily dependent on the interpreter's expertise and experience, resulting in strong subjectivity [7]. These approaches require exceptional operator skills while being inefficient, labor intensive, and accuracy-limited. In recent years, the rapid advancement of big data and artificial intelligence has prompted growing adoption of data mining and machine learning approaches for lithofacies identification research [8,9]. Support Vector Machines (SVM) demonstrate exceptional classification capabilities, with their application to sandstone reservoir lithology classification dating back to 2010 when A. Al-Anazi and colleagues pioneered its implementation in this field [10]. Building on this foundation, Bhattacharya and colleagues (2016) extended SVM application to shale reservoir lithofacies identification, with their superior classification accuracy validating SVM as both effective and practical for complex lithofacies analysis [11]. Liu et al. (2020) developed the Local Deep Multiple Kernel Learning SVM (LDMKL-SVM), which enhances computational accuracy through automated kernel function parameter learning [12]. Subsequently, Gao et al. (2022) achieved effective simplification of complex nonlinear problems by integrating triaxial vibration signal hybrid-domain features with SVM [13]. While SVM and other single-model approaches achieve acceptable classification accuracy, their poor generalizability becomes particularly apparent in complex, heterogeneous reservoirs where feature distributions vary substantially [14]. Consequently, researchers have increasingly turned their attention to ensemble decision tree algorithms, which offer stronger interpretability, lower sample size requirements, and enhanced robustness for lithofacies classification. Banerjee et al. (2024) successfully applied RF technology to lithofacies classification in India's Bokaro coal-field using wireline log data [15]. In a parallel methodological advancement, Wang et al. (2020) developed an enhanced lithology identification approach by integrating Hidden Markov Models with RF [16]. Merembayev et al. (2021) developed a lithofacies identification model using eXtreme Gradient Boosting (XGBoost), which significantly improved prediction accuracy through second-order derivative optimization of the loss function, with practical applications demonstrating the algorithm's reliability [17]. To further optimize computational efficiency, Light Gradient Boosting Machine (LightGBM) was subsequently developed by incorporating histogram-based techniques and leaf-wise growth strategies into the XGBoost framework [18], achieving reduced memory usage while maintaining high training speed. Liu et al. (2023) achieved high-precision lithofacies identification in southern Sichuan deep shale formations by integrating feature derivation with LightGBM modeling [19]. However, such ensemble models exhibit elevated complexity, including intricate hyperparameter tuning requirements

and substantially increased computational/storage costs, particularly demonstrating limited processing capacity when handling large-scale, high-dimensional datasets.

Neural network algorithms represent a powerful computational tool that simulates biological neuronal mechanisms to solve complex problems. Particularly renowned for their universal approximation theorem, these algorithms demonstrate exceptional capability in processing high-dimensional, nonlinear datasets [20]. Luo (2018) successfully applied Backpropagation Neural Networks (BPNN) to lithology prediction in continental shale oil reservoirs, achieving accurate classification of mudstone, calcareous mudstone, calcareous oil-bearing mudstone, and calcareous argillaceous oil shale [21,22]. Building on this foundation, Ameer-Zaimeche et al. [23] developed a Multilayer Perceptron Neural Network (MLPNN) by integrating multilayer perceptrons into a neural network framework. Concurrently, Feng et al. (2021) incorporated Markov transition matrices with Bayesian algorithms to establish a Bayesian Neural Network (B-ANN) [24]. Both advanced architectures demonstrated superior performance in lithofacies identification. With the rapid advancement of artificial intelligence, neural network algorithms have progressively evolved into diverse deep learning architectures tailored for different application scenarios. Given the continuous nature of well-logging data, researchers have employed Long Short-Term Memory (LSTM) networks to establish lithology identification models for complex reservoirs [25,26]. Notably, Li et al. (2021) developed an integrated CNN-LSTM framework that enhances feature recognition and extraction capabilities, achieving successful application in complex reservoir lithology classification [27].

Current intelligent lithofacies identification models primarily adopt three approaches: single-model architectures, homogeneous ensembles, and concatenated fusion. While these methods have improved recognition performance through optimized feature extraction and decision-making mechanisms, they still face critical challenges, including high computational complexity and limited interpretability. To address these issues, this paper proposes a novel hybrid super-fusion framework, named SRLCL. The SRLCL tackles limited lithofacies data and class imbalance via Newton-Weighted Oversampling enhances optimization efficiency with the Polar Lights Optimization (PLO) algorithm and overcomes feature extraction limitations by integrating five heterogeneous algorithms (SVM, RF, LightGBM, CNN, and LSTM) within a Stacking ensemble framework. This integration leverages their synergistic complementarity, achieving robust performance in lithofacies identification for deep shale gas reservoirs in the Luzhou block.

2. Geological Setting

The Sichuan Basin serves as a core area for shale gas exploration and development in China, having established several national shale gas demonstration zones—including Weiyuan, Changning, Fuling, and Zhaotong—primarily focused on mid-shallow shale gas resources [28]. Recent exploration studies indicate significant production potential in deep shale gas, with estimated geological resources of approximately $24.28 \times 10^{12} \text{ m}^3$. This enrichment characteristic is largely governed by the region's unique geological setting [29]. The basin exhibits a rhomboidal shape, stretching approximately 380–430 km from east to west and 310–330 km from north to south, covering an area of about $4 \times 10^4 \text{ km}^2$. It is situated in a relatively stable section of the Yangtze Block in southern China, surrounded by the Daba Mountains, Huaying Mountains, and Yunnan–Guizhou Plateau. Topographically, the basin is characterized by low elevations ranging from 250 to 750 m above sea level, with a general southeastward tilt (Figure 1).

The sample data were collected from the LZ block in the southern Sichuan Basin. The dataset comprises lithofacies types and well-logging data, with five lithofacies categories serving as output labels: (I) calcareous siliceous shale, (II) argillaceous siliceous shale,

(III) calcareous mixed shale, (IV) argillaceous mixed shale, and (V) mixed shale. The input features consist of well-logging curves, including acoustic travel time (AC), natural gamma ray (GR), bulk density (DEN), neutron porosity (CNL), deep resistivity (RD), flushed zone resistivity (RXO), as well as potassium (K), uranium (U), and thorium (Th) concentrations. The sample data were divided into a training set and a test set in a 7:3 ratio, and a 5-fold cross-validation method was employed, ensuring that all data were used for training and thus validating the robustness of the results. The original sample data are presented in Table 1.

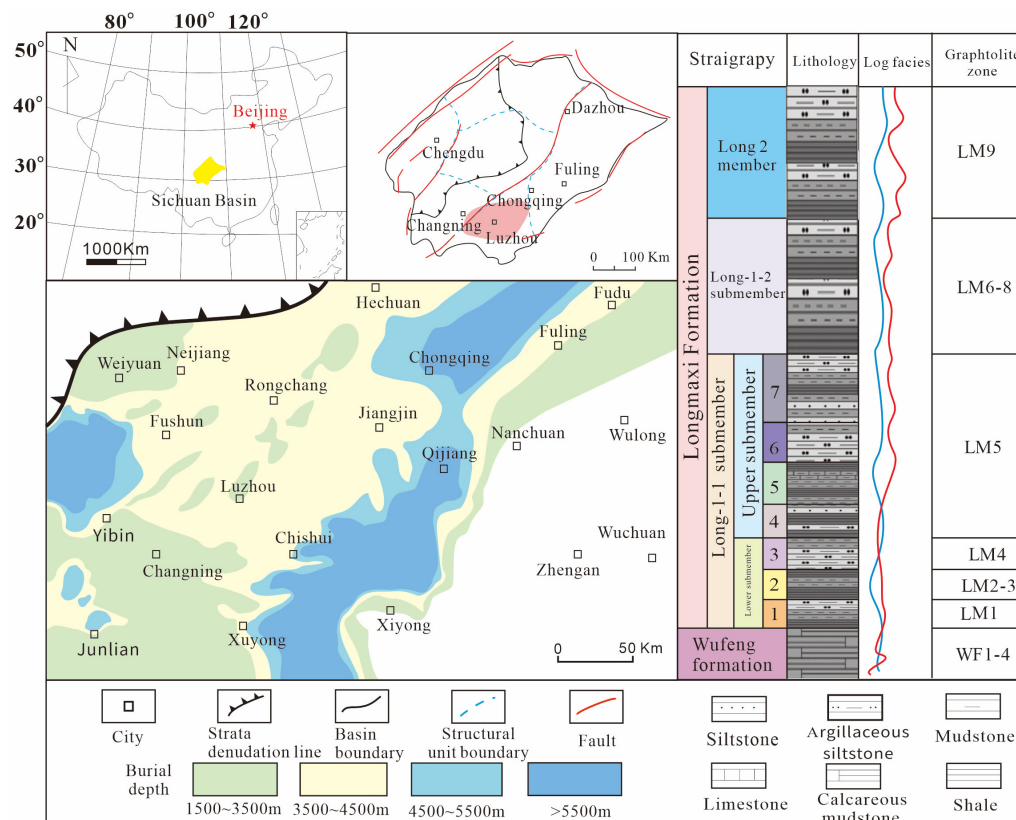


Figure 1. Location and structural schematic map of the study area.

Table 1. Presentation of partial original sample data.

[illegible]

3. Methodology

3.1. Polar Lights Optimizer

A novel metaheuristic algorithm named the Polar Lights Optimizer (PLO) is introduced [30]. The optimization process of PLO is primarily inspired by the physical phenomenon of auroras, simulating the motion trajectories of high-energy particles within the solar wind. The mathematical model of PLO incorporates innovative elements such as the Lorentz force, Newton's second law, a damping factor, and adaptive weighting. By designing strategies including rotational motion, auroral elliptical steps, dynamic step size, and particle collision, the algorithm mimics the movement of charged particles in Earth's magnetic field and atmosphere. These strategies are employed to enhance local search, global search, and avoidance of local optima, respectively, thereby achieving a dynamic balance between global exploration and local exploitation, and improving the effectiveness and precision of the optimization algorithm (Figure 2).

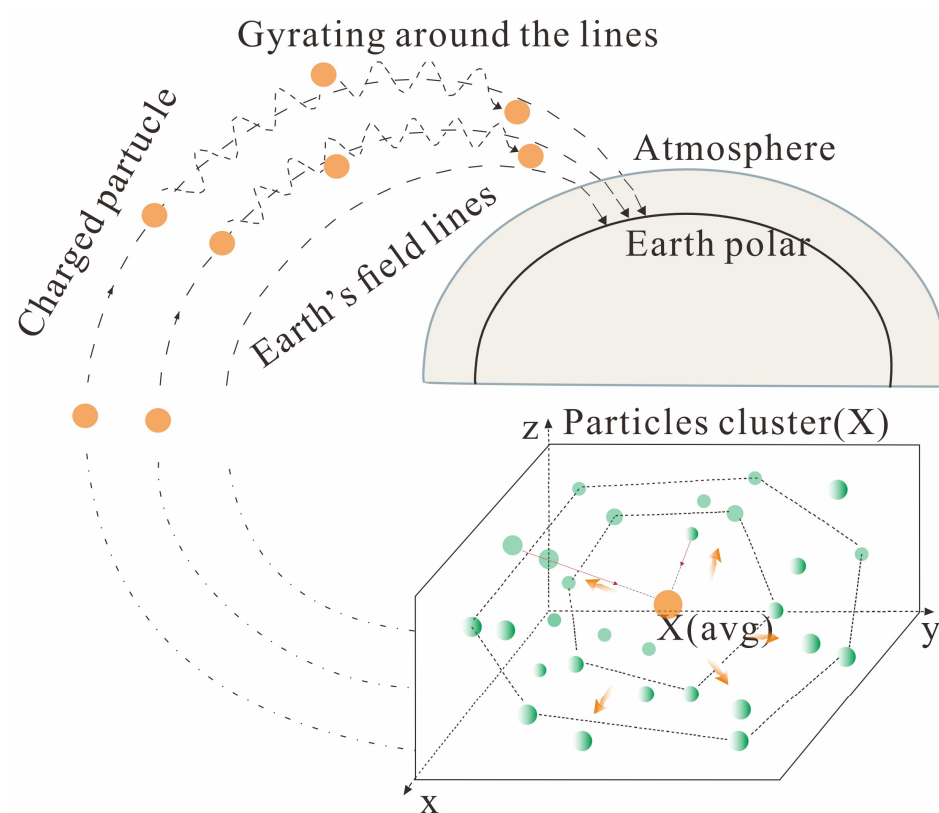


Figure 2. Schematic diagram of hyperparameter search in the Polar Light Optimizer.

The Polar Light Optimizer (PLO) incorporates a unique rotational motion model that simulates the deflection behavior of high-energy particles under Earth's magnetic field and their collision processes with atmospheric molecules. By emulating the velocity attenuation effect of the atmosphere on charged particles through physical laws, this model enables a more intensive search for optimal solutions during the local exploitation and progressively enhances convergence precision throughout iterations. Additionally, the algorithm introduces an auroral elliptical walking strategy, mimicking the phenomenon where high-energy particles gradually converge over polar regions to form luminous elliptical rings. Particles can move freely within these elliptical rings while chaotically following variations in the Earth's magnetic field. This dual behavior allows them to explore the entire solution space with dynamically adjusted step sizes. Guided by the

centroid of the particle swarm, this strategy ensures a coherent convergence direction and strengthens global exploration capability.

By integrating both rotational motion and auroral elliptical walking mechanisms, PLO achieves an effective balance between global exploration and local exploitation. Through dynamic adjustment of search step sizes and exploration strategies, the algorithm extensively explores potential optimal regions in the initial phases and increasingly refines its local search as iterations proceed. Notably, unlike many algorithms that require predefined additional parameters, PLO operates without the need for extra parameter settings. This characteristic offers significant advantages when addressing complex problems, simplifies practical application, and reduces dependence on problem-specific parameter tuning.

$$A_0 = Levy(d) \times (X_{avg}(j) - X(i, j)) + LB + r_1 \times (UB - LB) / 2 \quad (1)$$

$$X_{avg} = \frac{1}{N} \times \sum_{i=1}^N X(i) \quad (2)$$

$$X_{new}(i, j) = X(i, j) + r_2 \times (W_1 \times v(t) + W_2 \times A_0) \quad (3)$$

where A_0 represents the auroral elliptical walking trajectory, $Levy(d)$ denotes the step size value of Lévy flight, X_{avg} is the centroid position of the high-energy particle swarm which indicates the current position of a high-energy particle, UB and LB represent the upper and lower bounds of the solution space, r_1 and r_2 are random numbers within $[0, 1]$ simulating environmental disturbances, $X_{new}(i, j)$ corresponds to the updated position of the energy particle, $v(t)$ is the particle velocity, and W_1 and W_2 are weighting coefficients.

3.2. Newton-Weighted Oversampling Method

To address the issues of sparse data and extreme class imbalance in lithofacies classification of deep shale gas reservoirs, the Newton-Weighted Oversampling (NWO) method was employed to enhance the lithofacies dataset [31]. This approach effectively mitigates limitations of conventional oversampling techniques such as SMOTE and its variants, including noise generation, inadequate handling of boundary samples, and overfitting. The NWO method reduces synthetic noise and improves the distribution balance of minority classes by calculating minority class weights, eliminating feature noise, and generating feature subspaces, thereby enhancing the performance of the classification model.

To identify hard-to-learn samples and quantify the intensity of noise removal, weights are calculated based on the density of the initial sampling region and the average distance to the nearest majority class samples for each minority class instance. The weighting scheme adheres to the principles of density and proximity factors: higher weights are assigned to minority-class samples within sparse subconcepts to enhance their representation, and to those near the decision boundary to reduce their misclassification risk. The process begins by determining the neighborhood range to analyze its composition, followed by calculating weights to quantify each sample's boundary risk. In extreme cases, certain minority-class samples may account for over 90% of the total weight, resulting in only a very limited number of minority samples being selected for synthetic sample generation, and therefore, replacing these excessively large weights with high quantile values has minimal impact on the overall dataset generation. In normally distributed imbalanced datasets, however, the differences among the highest weights may be insignificant. To address this issue, assuming that max represents the maximum achievable value for $f(w_{min}^i)$, excessively large weights can be replaced with their quantile values.

The weight calculation formula is given as

$$w_{min}^i = C_{min}^i \times \frac{1}{D_{min}^i} \quad (4)$$

$$f(w_{min}^i) = \begin{cases} w_{min}^i & \text{if } w_{min}^i < \max \\ \max & \text{if } w_{min}^i > \max \end{cases} \quad (5)$$

In the formula, w_{min}^i represents the weight of the minority class, C_{min}^i denotes the proximity factor, and D_{min}^i stands for the density factor.

Prior to oversampling, feature noise is eliminated to prevent the exacerbation of distribution overlap. High-heat minority class samples are selected as heat sources based on their initial heat values, with their thermal energy diffusing in a hyperspheric pattern according to Newton's law of cooling. During diffusion, if majority class samples are present within the hypersphere, they are relocated outside. This process continues until the residual heat decays to a predefined equilibrium threshold, thereby effectively separating class-overlapping regions and providing a secure data distribution for subsequent oversampling. By defining an equilibrium heat parameter to terminate the noise removal process, the residual heat from minority class sample x_{min}^i to the j -th nearest majority-class sample is calculated as follows:

$$heat_m(x_{min}^i, x_{maj}^j) = G_num^i \cdot e^{-a(dis_max[i,j]-r^i)} \quad (6)$$

$$h_percent = \frac{S_{min}}{S_{maj}} \quad (7)$$

where x_{min}^i represents the minority class samples, G_num^i is the initial heat of the minority class samples, and α is the heat decay coefficient (its optimal values were determined using an optimization method, with the value ranges established based on empirical knowledge). dis_max is the distance matrix from the minority class to the majority class, r is the thermal diffusion radius, S_{min} is the number of minority class samples, and S_{maj} is the number of majority class samples.

By calculating minority class sample weights and eliminating feature noise, the weights and sampling regions for synthetic sample generation are determined. Subsequently, new minority class samples are generated based on a feature subspace generation mechanism. Figure 3 visually illustrates this sample generation process. For a given minority-class sample A, two neighbors are randomly selected from its k_{min} nearest neighbors, denoted as B and C. The direction vector $\overrightarrow{AD}$ is defined as follows, where both ε and β are random values within the range $[-1, 1]$. Once the direction vector for sample generation is determined, a new sample is created at a randomly selected position along this vector. The specific generation process is described by the following formula:

$$\overrightarrow{AD} = \varepsilon \overrightarrow{AB} + \beta \overrightarrow{AC} \quad (8)$$

$$x_{gen} = x_{min}^i + \mu * r^i * \frac{\overrightarrow{AD}}{\|\overrightarrow{AD}\|} \quad (9)$$

where μ is a random number in the range $[0, 1]$, indicating that the newly synthesized minority class sample lies within a hypersphere centered at x_{min}^i with radius r^i .

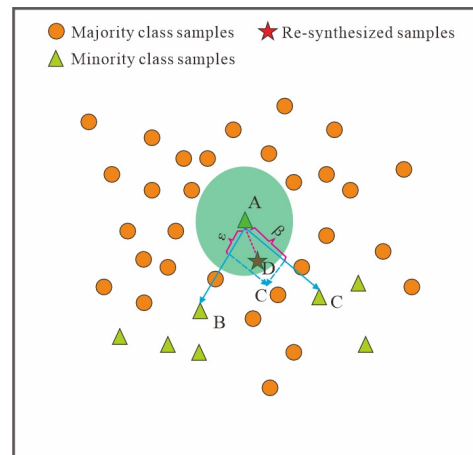


Figure 3. Schematic diagram of the feature subspace generation mechanism.

3.3. Heterogeneous Fusion Model

3.3.1. Stacking Ensemble Learning Framework

Stacking is an ensemble method that combines multiple heterogeneous base learners through a meta-learner [32]. Unlike classical Bagging (which uses parallel voting of homogeneous learners) and Boosting (which employs sequential weighting of homogeneous learners), Stacking exhibits two core characteristics: heterogeneity of base learners and learnability of their combinations. The Stacking algorithm therefore generally adopts a two-layer architecture consisting of base learners and a meta-learner, along with the dual-layer learning mechanism. The first layer comprises diverse types of heterogeneous machine learning models that learn from raw sample data to generate primary predictive features. The second layer consists of a meta-learner, which performs secondary learning based on the primary predictive features from the first layer to achieve global optimization, thereby systematically reducing model bias and variance. This structure enables adaptive integration of the discriminative capabilities of different models, achieving complementary advantages and ultimately enhancing the overall generalization performance and predictive accuracy of the model (Figure 4).

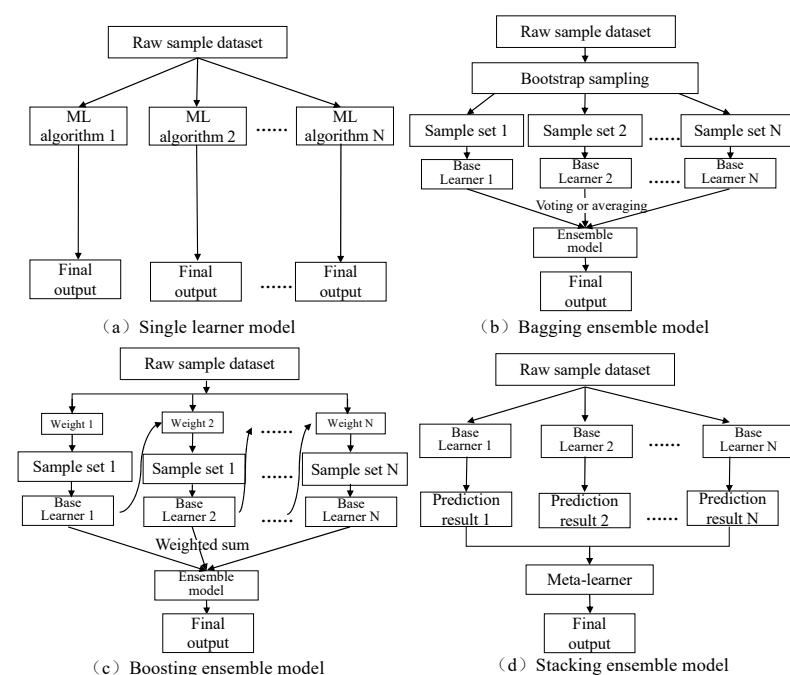


Figure 4. Comparison of single model and different ensemble modeling frameworks.

3.3.2. Intelligent Lithofacies Identification Heterogeneous Fusion Model

The Stacking framework offers considerable flexibility in the selection and combination of diverse base learners. It can integrate not only conventional simple models but also ensemble models as base learners for higher-order integration and even achieve cross-level integration of single models, ensemble models, and deep learning architectures. This study therefore selects five structurally distinct algorithms as base learners—Support Vector Machine (SVM), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM)—in order to fully leverage the distinct advantages of existing machine learning algorithms and thoroughly extract deep-level latent lithofacies information from well-logging data. Note that for CNN and LSTM, a sliding window is used to construct samples along the depth sequence. Within each window, multiple logging curves are treated as input channels, while the depth sequence itself is regarded as time steps and fed into the model. A Multilayer Perceptron (MLP) is then employed as the meta-learner to construct the heterogeneous fusion model (SRLCL), integrating single models, ensemble models, and deep learning models into a unified framework for intelligent lithofacies identification in deep shale gas reservoirs.

The framework of the SRLCL model and its lithofacies identification workflow are illustrated in Figure 5. The training process comprises two stages: base-learner training and meta-learner training, as detailed below. First, a balanced and high-quality lithofacies sample set is constructed through data preprocessing and NWO. This is performed to mitigate the effects of data noise and class imbalance. The resulting dataset is then randomly split into training and testing sets in a 7:3 ratio. Each of the five base learners is independently trained using k-fold cross-validation, ensuring that every data sample participates in both training and validation to minimize partitioning bias. Subsequently, the PLO algorithm is applied to optimize the hyperparameters of the models, generating optimized preliminary lithofacies predictions. Finally, the initial predictions from all base learners are used as input features to train the meta-learner. The PLO algorithm is once again employed to optimize the meta-learner, yielding the final lithofacies identification results.

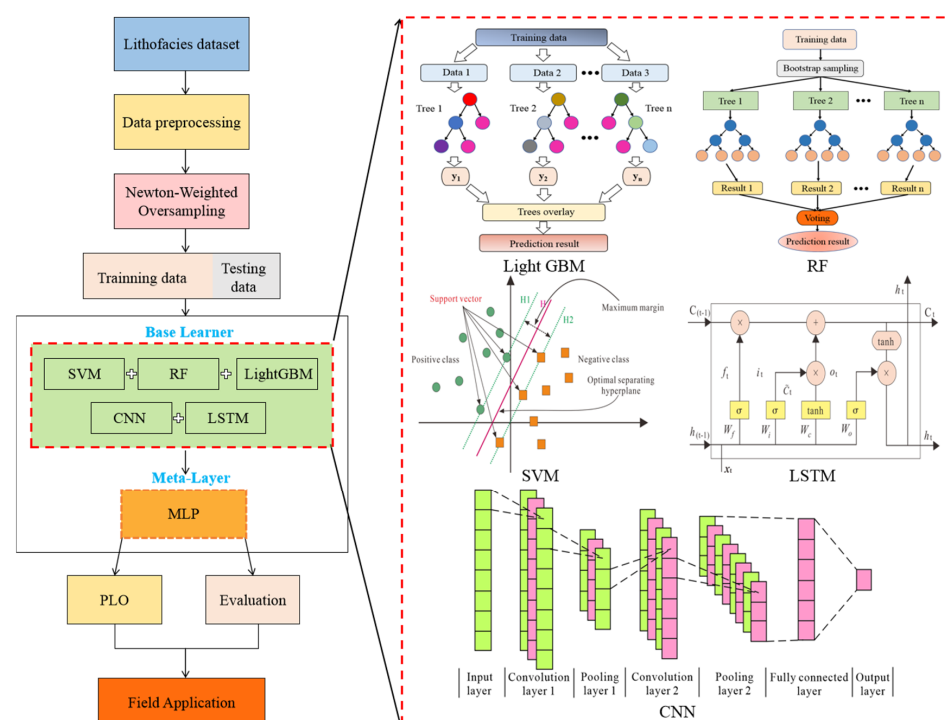


Figure 5. Schematic diagram of the SRLCL framework and workflow.

4. Results and Discussion

4.1. Model Performance

The fusion algorithm was trained and tested on the NWO-resampled dataset, with model optimization performed using the MSICS algorithm. Table 2 presents the optimal hyperparameter configurations for each base learner within the fusion framework.

Table 2. Statistics of optimal hyperparameter configurations for base learners.

Algorithm	Hyperparameter	Best Value
SVM	C	4.29
	kernel	rbf
	gamma	scale
	degree	4
RF	max_depth	40
	min_samples_split	2
	min_samples_leaf	2
	n_estimators	125
	criterion	gini
LightGBM	n_estimators	32
	learning_rate	0.09
	max_depth	20
	num_leaves	68
CNN	epochs	10
	batch_size	20
	filters	48
	kernel_size	4
	dropout_rate	0.6
LSTM	epochs	7
	batch_size	21
	units	167
	dropout_rate	0.33

Lithofacies identification was conducted using the SRLCL model with optimal hyperparameter configurations. The classification results for the five lithofacies types are presented via a confusion matrix (Figure 6). The diagonal values of the matrix represent the number or probability of correctly classified samples for each category. Visual analysis of the confusion matrix indicates that the SRLCL model achieves generally favorable performance in identifying the five lithofacies types. Among these, the argillaceous siliceous shale facies shows relatively lower recognition accuracy, with approximately 18% of its samples misclassified. The other four lithofacies types demonstrate excellent identification results: 98% of both calcareous siliceous shale and calcareous mixed shale facies are correctly classified, with only two misclassified samples each. The argillaceous mixed shale and mixed shale facies perform most prominently, achieving perfect classification accuracy.

To further evaluate performance, key metrics including Precision, Recall, and F1-Score were calculated based on the confusion matrix. Table 3 and Figure 7 present the statistical results of these classification evaluation metrics for the five lithofacies types using the fusion model. Analysis revealed that the SRLCL model performed well in identifying all lithofacies types, achieving consistently high scores in Precision, Recall, and F1-Score. For the five lithofacies, the precision scores were 0.93, 1.00, 0.99, 0.93, and 0.83, respectively, with an average of 0.94. The recall rates were 0.98, 0.72, 0.98, 1.00, and 1.00, yielding an average of 0.94. The F1-scores were 0.95, 0.84, 0.98, 0.96, and 0.91, resulting in an average of 0.93. The overall classification accuracy of the model reached 93%. Although a minor portion of

argillaceous siliceous shale was prone to misclassification, the other four lithofacies types were accurately identified, demonstrating the excellent lithofacies identification capability of the SRLCL model.

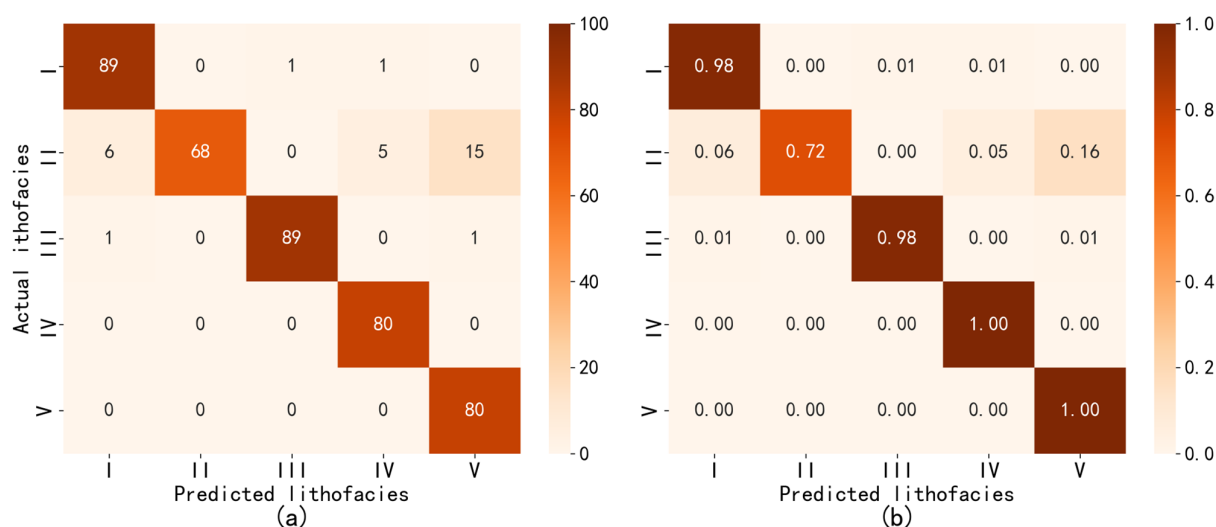


Figure 6. Confusion matrix of lithofacies identification results: (a) Original confusion matrix; (b) normalized confusion matrix.

Table 3. Statistical table of evaluation metrics for lithofacies identification results.

Lithofacies Types	Code Name	Precision	Recall	F1-Score	Accuracy
Calcareous siliceous shale	I	0.93	0.98	0.95	0.93
Argillaceous siliceous shale	II	1.00	0.72	0.84	
Calcareous mixed shale	III	0.99	0.98	0.98	
Argillaceous mixed shale	IV	0.93	1.00	0.96	
Mixed shale	V	0.83	1.00	0.91	
Macro-average		0.94	0.94	0.93	

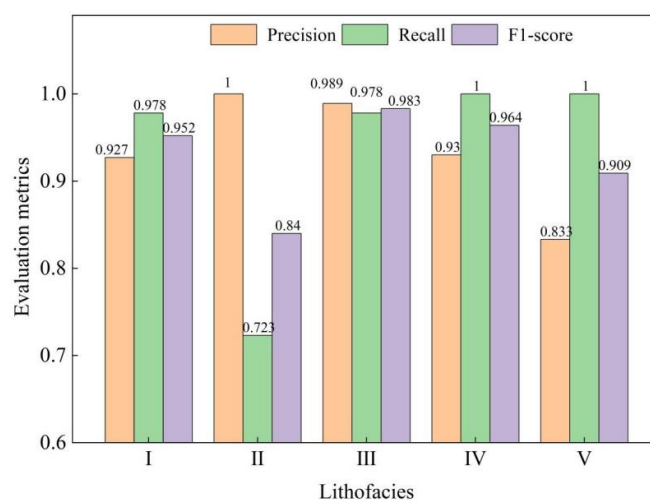


Figure 7. Histogram of lithofacies identification evaluation metrics.

4.2. Comparative Experiments

4.2.1. Performance Comparison of SRLCL with Different Sampling Methods

Figures 8 and 9 visually present the corresponding lithofacies identification results. Analysis reveals that without resampling, the model's F1-scores are highly polarized across

different lithofacies. The calcareous siliceous shale facies achieves an F1-score of 0.86, while the argillaceous mixed shale facies attains only 0.06. This phenomenon indicates that under extreme class imbalance, the learning model tends to overemphasize majority class samples such as calcareous siliceous shale while neglecting minority class samples like mixed shale. Due to insufficient discriminative features extracted from minority class samples, the boundaries between lithofacies categories become blurred, leading to classification outcomes with strong randomness and variability.

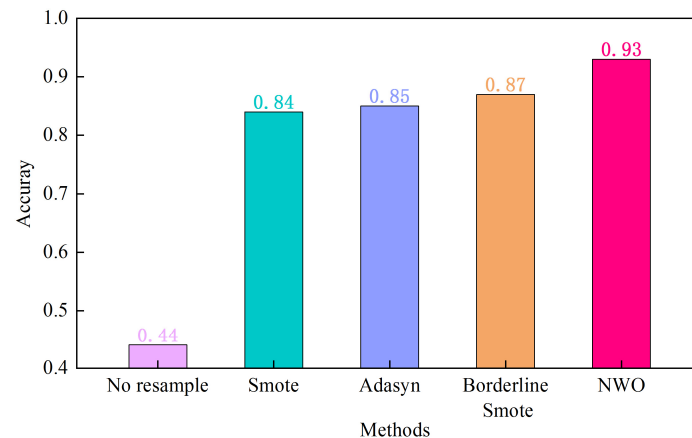


Figure 8. Histogram of accuracy rates for SRLCL under different sampling techniques.

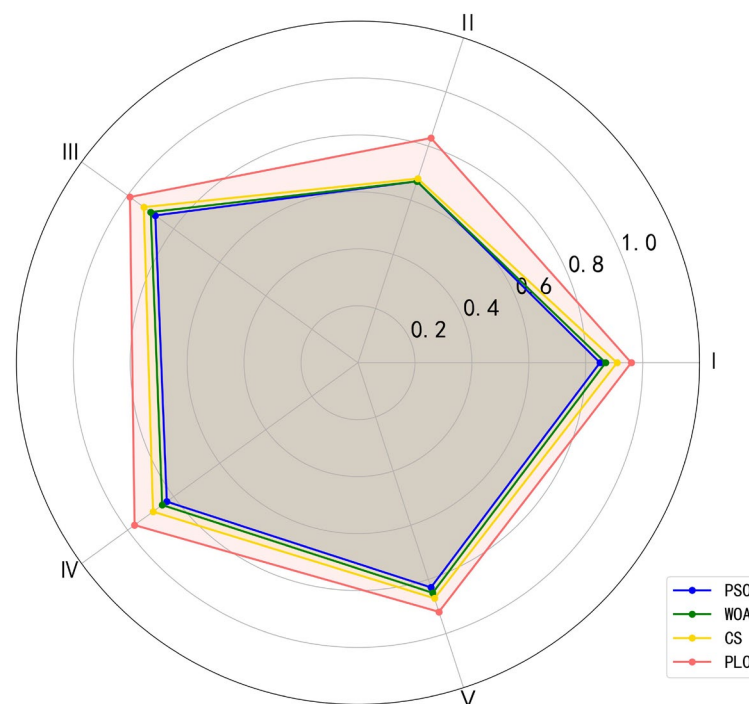


Figure 9. Radar chart of F1-scores for the SRLCL model under different resampling techniques.

After applying four oversampling methods (NWO, Borderline-SMOTE, ADASYN, and SMOTE) to resample the original dataset, the lithofacies identification performance of the SRLCL model improved significantly, with accuracy exceeding 0.84 and macro-average F1-score surpassing 0.83. Both metrics increased by over 40% compared to the model without resampling. Notably, the NWO method proposed in this study demonstrated the most effective class imbalance handling capability, outperforming the suboptimal Borderline-SMOTE method by improving accuracy by 6% and the macro-average F1-score by 7%.

4.2.2. Performance Comparison of SRLCL with Different Optimization Algorithms

As shown in Figures 10 and 11, among the three swarm intelligence optimization algorithms (PSO, WOA, and CS) evaluated, the CS-optimized SRLCL model delivered the best lithofacies identification performance, as shown in Figures 10 and 11, achieving an overall accuracy of 0.86 and a macro-average F1-score of 0.85. Among the five lithofacies types, four exhibited F1-scores approaching 0.9, while the argillaceous siliceous shale facies showed a relatively lower F1-score. Although WOA and PSO produced less competitive results, both still achieved accuracy and macro-average F1-scores above 0.8.

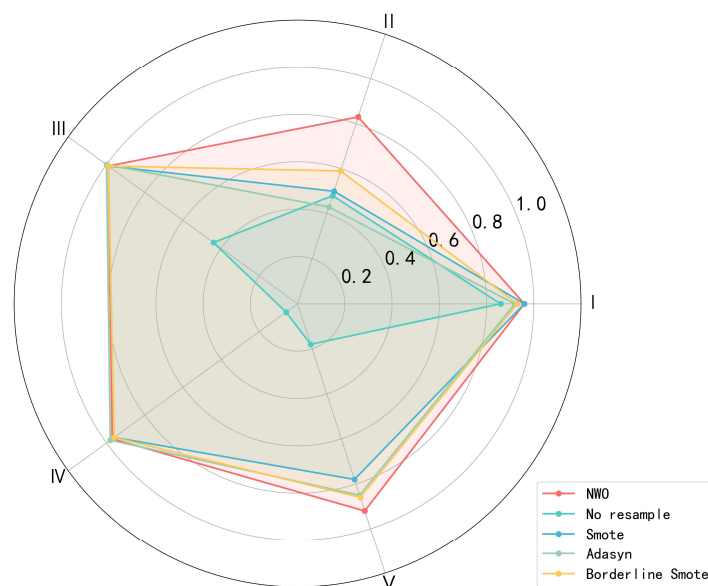


Figure 10. Comparative radar chart of F1-scores for the SRLCL model with different optimization algorithms.

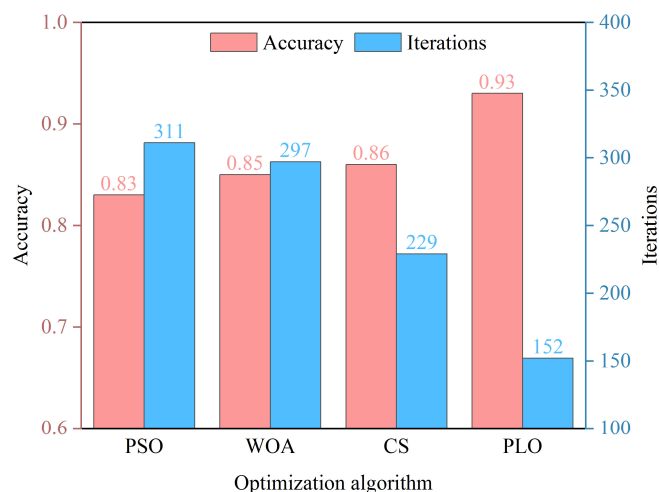


Figure 11. Comparison of iteration counts for optimal lithofacies identification results using different optimization algorithms.

The SRLCL model optimized with the PLO algorithm demonstrates substantial improvement in lithofacies identification performance. Compared with the CS algorithm, the model's accuracy increases from 0.86 to 0.93, and its macro-average F1-score rises from 0.85 to 0.93, representing an improvement of over 7% in both metrics. The most notable enhancement is observed for the argillaceous siliceous shale facies, whose F1-score improves from 0.67 to 0.84—an increase of approximately 17%. Furthermore, PLO exhibits

high convergence efficiency, requiring only 152 iterations to determine the optimal hyper-parameter combination for the SRLCL model and obtain the global optimum. Compared to PSO (311 iterations), WOA (297 iterations), and CS (229 iterations), the iteration counts are reduced by 51.1%, 48.8%, and 33.6%, respectively, indicating a significant improvement in optimization efficiency.

4.2.3. Interpretability Analysis of SRLCL Model

Figure 12 presents the input feature importance alongside SHAP beeswarm plots for the SRLCL model. This combined visualization illustrates both the global significance of each feature and their sample-wise contributions. The well-logging curves are ranked by contribution as follows: AC > GR > CNL > DEN > Th > LnRT > K > U. AC is the most influential feature, followed by GR, CNL, and DEN—a finding consistent with established principles of petrophysical interpretation. For the top five features, low-value samples (blue dots) are predominantly located in the negative SHAP region, indicating that lower values exert a stronger negative influence on predictions. The sparse distribution and short horizontal spread of these dots indicate both a limited proportion of low-value samples and their restrained inhibitory effect. Mitigating the influence of these specific samples could potentially enhance model performance. In contrast, high-value samples (red dots) are densely concentrated in the positive SHAP region, with extensive horizontal spread and noticeable tails, demonstrating their pronounced positive influence on predictions.

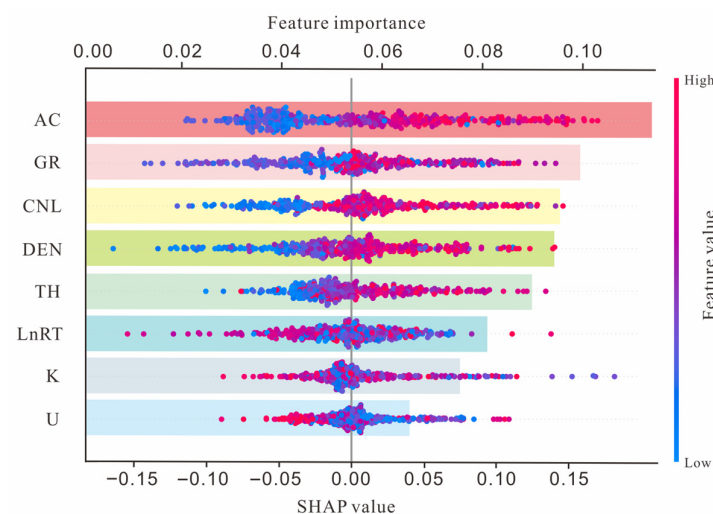


Figure 12. Feature importance and SHAP visualization of the SRLCL model.

Features such as LnRT, K, and U, however, show dots clustered near zero, with mixed red and blue points and blurred boundaries. This pattern implies a weak overall influence and no clear linear relationship with prediction outcomes. Nevertheless, these features may participate in complex nonlinear associations with lithofacies and still play roles in local decision-making.

Figure 13 illustrates the contribution levels of the five base learners to the SRLCL model. LightGBM demonstrates the highest contribution, followed by Random Forest (RF), with both scoring above 2.5. By leveraging their strong feature selection capabilities, they effectively capture nonlinear relationships and high-order interactions, thereby serving as the model's core pillars. CNN and SVM both score 1.5; although relatively lower, their roles remain indispensable: CNN enhances spatial feature extraction and local pattern recognition, while SVM optimizes classification boundaries. LSTM scores below 1, indicating relatively weaker performance in this multi-class task, yet it provides unique

advantages in capturing sequential features. By complementing the strengths of other models, it contributes to enhanced overall predictive performance.

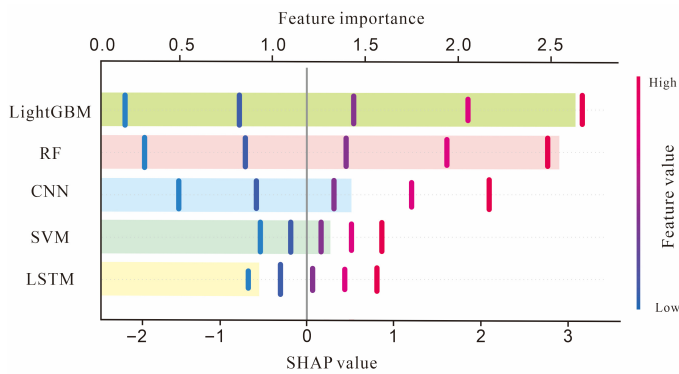


Figure 13. SHAP-based visualization of base model contributions.

4.3. Case Validation and Application

4.3.1. Test Well Validation

To validate the generalization capability of the constructed SRLCL model, lithofacies identification was performed on the target interval of Well W1. The evaluation metrics (Table 4) demonstrate that the SRLCL model achieved an identification accuracy of 0.91 and a macro-average F1-score of 0.91 on the test well, with other evaluation metrics also exceeding 0.9, indicating excellent performance. Furthermore, the single-well log (Figure 14) reveals strong consistency between core-derived lithofacies and identified lithofacies along the well profile, with an agreement rate exceeding 90%. This validation case thus confirms the model’s robust discriminatory power and generalization capability, demonstrating its practical reliability for field applications.

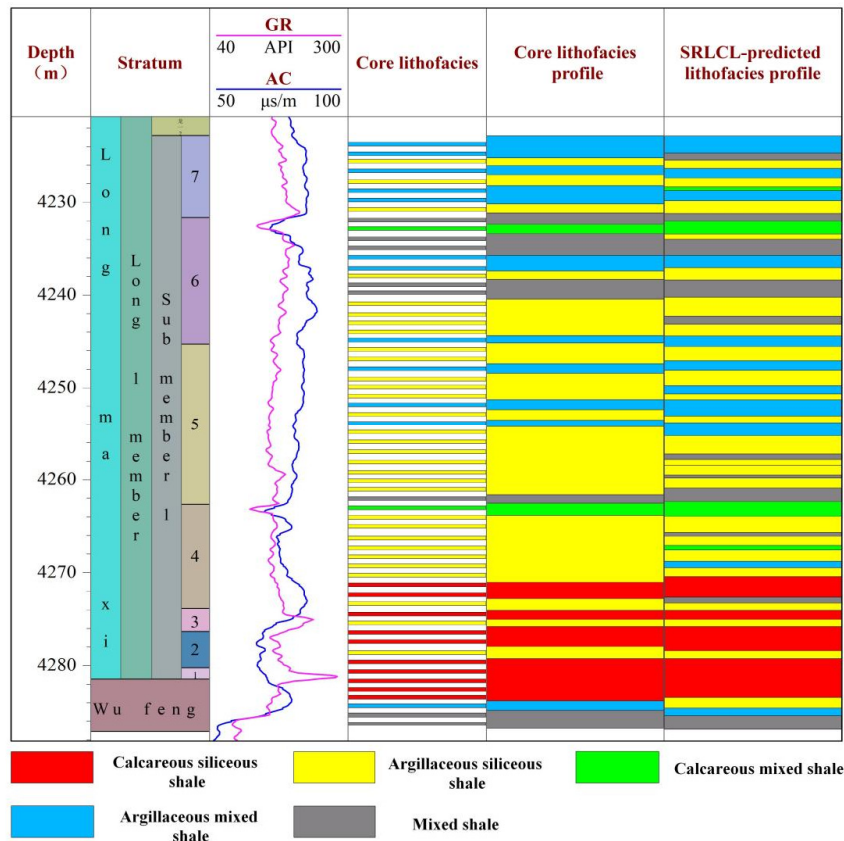


Figure 14. Comparative single-well log of identified versus actual lithofacies profiles for Well X13.

Table 4. Statistics of evaluation metrics for lithofacies identification in the test well.

Lithofacies Types	Precision	Recall	F1-Score	Accuracy
Calcareous siliceous shale	0.96	0.95	0.98	0.91
Argillaceous siliceous shale	0.93	0.73	0.75	
Calcareous mixed shale	0.99	0.99	0.99	
Argillaceous mixed shale	0.89	0.97	0.93	
Mixed shale	0.82	0.96	0.88	
Macro-average	0.92	0.92	0.91	

4.3.2. Production Performance Simulation Based on Geological Model

A production dynamics validation case was conducted for Well LX-H1 in the X block using the embedded discrete fracture model. The horizontal section of this well is located in Sub-layer 1 of Member 1 of the Longyi Formation, with a length of 2300 m. The well underwent 267-cluster fracturing across 30 stages and has been producing for over 1100 days, achieving a cumulative gas production of $7418.28 \times 10^4 \text{ m}^3$. During the initial constant-rate production phase, gas productivity was primarily contributed by the stimulated reservoir volume (SRV) region. After one year, as production declined, the flow mechanism shifted to matrix-to-fracture supply. Subsequently, under controlled-pressure production, the output stabilized, achieving a dynamic balance between matrix and fracture systems.

Two three-dimensional geological models were deployed for Well LX-H1 to perform history matching and cumulative gas production prediction: (a) a conventional sequential Gaussian simulation model and (b) a lithofacies-controlled model based on intelligent lithofacies identification. While both models exhibit consistent macroscopic trends, they show significant local differences. The conventional model demonstrates greater spatial continuity and homogeneity with smooth parameter gradients, whereas the proposed model exhibits distinct zoning characteristics with clear boundary transitions. As shown in Figure 15, the conventional model yields an average fracture length of 205 m and a single-stage SRV of 98.25 m^3 . In contrast, the proposed model yields fracture lengths predominantly distributed around 220 m and a single-stage SRV of 105.16 m^3 , representing an approximately 8% improvement over the conventional approach.

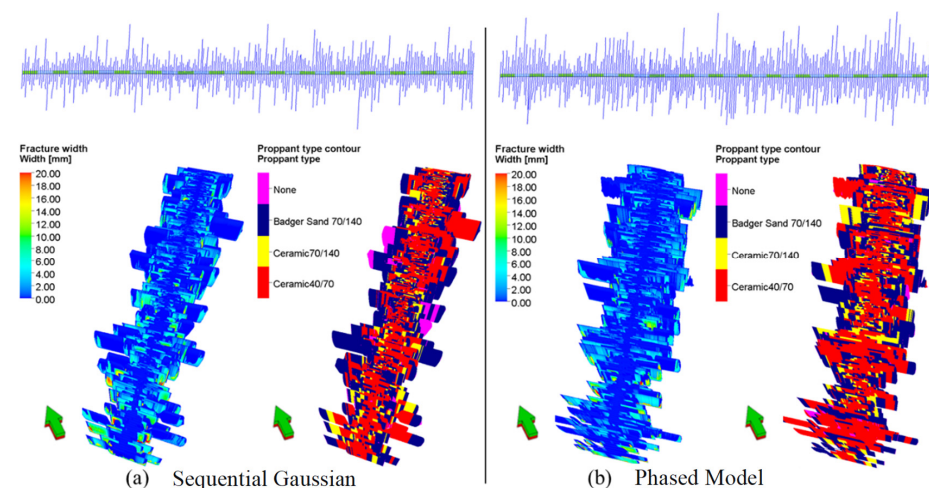


Figure 15. Hydraulic fracture network morphology of Well W1 under different geological models ((a) conventional sequential Gaussian simulation model, (b) lithofacies-controlled model).

The production history matching results in Figures 16 and 17 show that Model ①, constructed using sequential Gaussian simulation, predicts a cumulative gas production of $5981.43 \times 10^8 \text{ m}^3$, significantly lower than the actual cumulative production, with a fitting error of 19.38%. In contrast, Model ②, built based on lithofacies control, predicts

a cumulative gas production of $6959.99 \times 10^4 \text{ m}^3$, with a fitting error of only 6.18%. This represents a 13.2% reduction in error compared to Model ① and indicates close alignment with actual production data, demonstrating that Model ② better reflects actual formation conditions. It should be noted that both models predicted lower production than actual values, as the simulation did not account for field enhancement and production stabilization measures implemented during operations. Furthermore, in the 15-year cumulative production forecast, the two models yielded predictions of $11.203 \times 10^8 \text{ m}^3$ and $1.402 \times 10^8 \text{ m}^3$, respectively, confirming that the geological model developed using the proposed method can deliver better development benefits.

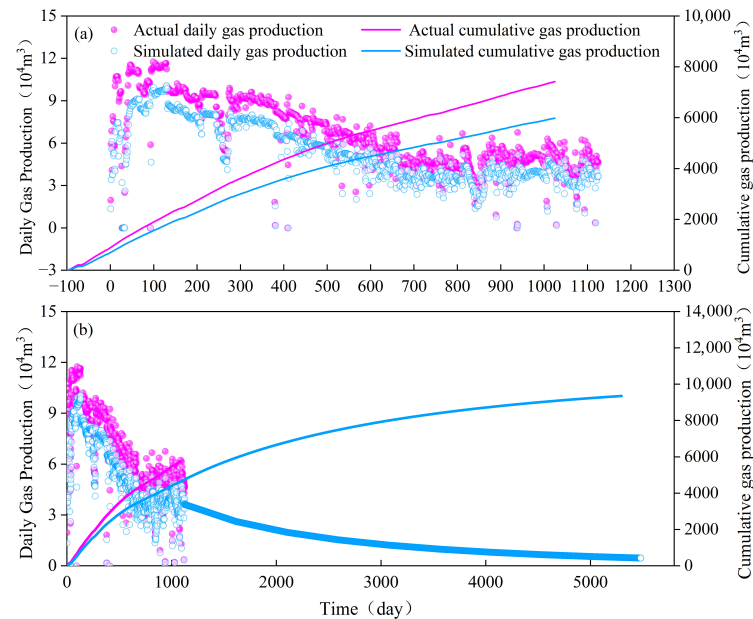


Figure 16. Production history matching and cumulative gas production forecast of Model ① based on sequential Gaussian interpolation((a) is the history-matching of production performance, (b) is the future production forecast).

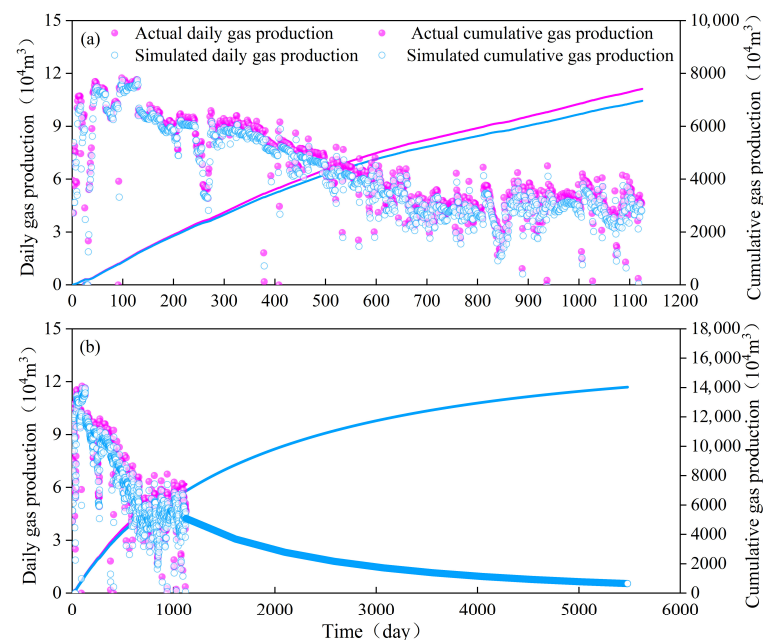


Figure 17. Production history matching and cumulative gas production forecast of Model ② based on the lithofacies-controlled modeling approach proposed in this study ((a) is the history-matching of production performance, (b) is the future production forecast).

5. Conclusions

By integrating improvements in optimization algorithms, data resampling methods, and intelligent algorithm fusion, a heterogeneous integrated model (SRLCL) for intelligent lithofacies identification in deep shale gas reservoirs has been developed. The following results and findings are summarized as follows:

- (1) The NWO method effectively addresses the issues of sparse data and class imbalance in deep shale gas reservoir lithofacies, increasing identification accuracy by over 40% compared to non-resampled data and by more than 6% compared to conventional resampling techniques. The PLO algorithm further optimizes model efficiency and accuracy, enabling the SRLCL model to achieve optimal performance after only 152 iterations—a 33.6% reduction in iteration count accompanied by a 7% improvement in accuracy.
- (2) Based on the Stacking ensemble learning framework, the SRLCL model overcomes the limitations of traditional single-model and homogeneous ensemble approaches, fully leveraging the complementary advantages of heterogeneous algorithms in characterizing multi-level features of well-logging data. The overall lithofacies identification accuracy of the SRLCL model reaches 93%.
- (3) Among the base learners in the SRLCL model, LightGBM and RF demonstrate outstanding feature selection and interaction capabilities, contributing the most; CNN and SVM follow, responsible for local feature extraction and classification boundary optimization, respectively; although LSTM underperforms overall, its unique ability to capture sequential features provides important supplementary value to the model.
- (4) Using lithofacies control to construct the geological model enables fine characterization and representation of deep shale reservoirs in the X block. Integrated with numerical simulation technology, the model facilitates integrated simulation of fractured horizontal well production in deep shale gas reservoirs. The predicted cumulative gas production shows an error of only 6.18% compared to actual data, indicating closer alignment with real formation conditions. Furthermore, the 15-year cumulative gas production forecast reaches $1.402 \times 10^8 \text{ m}^3$, demonstrating favorable development potential.

Author Contributions: Y.L.: Conceptualization, Writing (original draft), Resources, and Methodology; J.W.: Writing (review and editing), Validation, and Data curation; B.Z.: Supervision and Conceptualization; C.L.: Supervision and Project administration; F.D.: Visualization, Software, and Conceptualization; B.C.: Data curation and visualization; C.Y.: Visualization and Conceptualization; J.Y.: Visualization and Conceptualization; K.T.: Supervision and Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Science and Technology Special Project of China National Petroleum Corporation (CNPC) “Study on Shale Oil and Gas Enrichment Mechanisms and Reservoir Geomechanics Evaluation Technology” (Grant No. 2024DJ8706).

Data Availability Statement: Data will be made available on request.

Acknowledgments: We extend our sincere gratitude to the PetroChina Research Institute of Petroleum Exploration & Development and Chengdu University of Technology for their valuable support and assistance in terms of data resources and technical methodologies. We would also like to express our heartfelt thanks to all the institutions and individuals who have contributed to this research.

Conflicts of Interest: Authors Yi Liu, Boning Zhang, Bingyi Chen, Chen Yang, Jing Yang and Kai Tong were employed by Chengdu Northern Petroleum Exploration and Development Technology Co., Ltd. and Zhenhua Oil Co., Ltd. Author Jin Wu was employed by PetroChina Research Institute of Petroleum Exploration and Development. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Hota, R.N.; Maejima, W.; Maejima, W. Comparative Study of Cyclicity of Lithofacies in Lower Gondwana Formations of Talchir Basin, Orissa, India; A Statistical Analysis of Subsurface Logs. *Gondwana Res.* **2004**, *7*, 353–362. [\[CrossRef\]](#)
2. Lin, B.; Jin, Y.; Cao, Q.; Meng, H.; Pang, H.; Wei, S. Developing a Large Language Model for Oil- and Gas-Related Rock Mechanics: Progress and Challenges. *Nat. Gas Ind. B* **2025**, *12*, 110–122. [\[CrossRef\]](#)
3. Loucks, R.G.; Ruppel, S.C. Mississippian Barnett Shale: Lithofacies and Depositional Setting of a Deep-Water Shale-Gas Succession in the Fort Worth Basin, Texas. *AAPG Bull.* **2007**, *91*, 579–601. [\[CrossRef\]](#)
4. Zhao, L.; Xia, P.; Fu, Y.; Wang, K.; Mou, Y. Types of Lithofacies in the Lower Cambrian Marine Shale of the Northern Guizhou Region and Their Suitability for Shale Gas Exploration. *Nat. Gas Ind. B* **2024**, *11*, 469–481. [\[CrossRef\]](#)
5. Hao, F.; Zou, H.; Lu, Y. Mechanisms of Shale Gas Storage: Implications for Shale Gas Exploration in China. *AAPG Bull.* **2013**, *97*, 1325–1346. [\[CrossRef\]](#)
6. Zou, C.; Zhu, R.; Chen, Z.Q.; Ogg, J.G.; Wu, S.; Dong, D.; Qiu, Z.; Wang, Y.; Wang, L.; Lin, S.; et al. Organic-Matter-Rich Shales of China. *Earth-Sci. Rev.* **2019**, *189*, 51–78. [\[CrossRef\]](#)
7. Li, M.; Zhang, X.; Wang, K. Risk Prediction of Gas Hydrate Formation in Wellbores and Subsea Gathering Systems of Deep-Water Turbidite Reservoirs: Case Analysis from the South China Sea. *Reserv. Sci.* **2025**, *1*, 52–72. [\[CrossRef\]](#)
8. Wang, G.; Carr, T.R.; Ju, Y.; Li, C. Identifying Organic-Rich Marcellus Shale Lithofacies by Support Vector Machine Classifier in the Appalachian Basin. *Comput. Geosci.* **2014**, *64*, 52–60. [\[CrossRef\]](#)
9. Zheng, D.; Hou, M.; Chen, A.; Zhong, H.; Qi, Z.; Ren, Q.; You, J.; Wang, H.; Ma, C. Application of Machine Learning in the Identification of Fluvial-Lacustrine Lithofacies from Well Logs: A Case Study from Sichuan Basin, China. *J. Pet. Sci. Eng.* **2022**, *215*, 110610. [\[CrossRef\]](#)
10. Al-Anazi, A.; Gates, I.D. A Support Vector Machine Algorithm to Classify Lithofacies and Model Permeability in Heterogeneous Reservoirs. *Eng. Geol.* **2010**, *114*, 267–277. [\[CrossRef\]](#)
11. Bhattacharya, S.; Carr, T.R.; Pal, M. Comparison of Supervised and Unsupervised Approaches for Mudstone Lithofacies Classification: Case Studies from the Bakken and Mahantango-Marcellus Shale, USA. *J. Nat. Gas Sci. Eng.* **2016**, *33*, 1119–1133. [\[CrossRef\]](#)
12. Liu, X.Y.; Zhou, L.; Chen, X.H.; Li, J.Y. Lithofacies Identification Using Support Vector Machine Based on Local Deep Multi-Kernel Learning. *Pet. Sci.* **2020**, *17*, 954–966. [\[CrossRef\]](#)
13. Gao, K.; Jiao, S. Research on Lithology Identification Based on Multi-Sensor Hybrid Domain Information Fusion and Support Vector Machine. *Earth Sci. Inform.* **2022**, *15*, 1101–1113. [\[CrossRef\]](#)
14. Wu, J.; Ansari, U. From CO₂ Sequestration to Hydrogen Storage: Further Utilization of Depleted Gas Reservoirs. *Reserv. Sci.* **2025**, *1*, 19–35. [\[CrossRef\]](#)
15. Banerjee, A.; Mukherjee, B.; Sain, K. Machine Learning Assisted Model Based Petrographic Classification: A Case Study from Bokaro Coal Field. *Acta Geod. Geophys.* **2024**, *59*, 463–490. [\[CrossRef\]](#)
16. Wang, P.; Chen, X.; Wang, B.; Li, J.; Dai, H. An Improved Method for Lithology Identification Based on a Hidden Markov Model and Random Forests. *Geophysics* **2020**, *85*, IM27–IM36. [\[CrossRef\]](#)
17. Merembayev, T.; Kurmangaliyev, D.; Bekbauov, B.; Amanbek, Y. A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan. *Energies* **2021**, *14*, 1896. [\[CrossRef\]](#)
18. Gu, Y.; Zhang, D.; Bao, Z. Carbonate Lithofacies Identification Using an Improved Light Gradient Boosting Machine and Conventional Logs: A Demonstration Using Pre-Salt Lacustrine Reservoirs, Santos Basin. *Carbonates Evaporites* **2021**, *36*, 79. [\[CrossRef\]](#)
19. Liu, Y.; Zhu, R.; Zhai, S.; Li, N.; Li, C. Lithofacies Identification of Shale Formation Based on Mineral Content Regression Using LightGBM Algorithm: A Case Study in the Luzhou Block, South Sichuan Basin, China. *Energy Sci. Eng.* **2023**, *11*, 4256–4272. [\[CrossRef\]](#)
20. Qian, H.; Geng, Y.; Wang, H. Lithology Identification Based on Ramified Structure Model Using Generative Adversarial Network for Imbalanced Data. *Geoenergy Sci. Eng.* **2024**, *240*, 213036. [\[CrossRef\]](#)
21. Luo, H.; Lai, F.; Dong, Z. A Lithology Identification Method for Continental Shale Oil Reservoir Based on BP Neural Network. *J. Geophys. Eng.* **2018**, *15*, 895–908. [\[CrossRef\]](#)
22. Cao, L.; Lv, M.; Li, C.; Sun, Q.; Wu, M.; Xu, C.; Dou, J. Effects of Crosslinking Agents and Reservoir Conditions on Hydraulic Fracture Propagation in Coal Reservoirs. *Reserv. Sci.* **2025**, *1*, 36–51. [\[CrossRef\]](#)
23. Ameur-Zaimeche, O.; Zeddouri, A.; Heddam, S.; Boumezeur, A.; Bouguessa, S. Lithofacies Prediction in Non-Cored Wells from the Sif Fatima Oil Field (Berkine Basin, Southern Algeria): A Comparative Study of Multilayer Perceptron Neural Network and Cluster Analysis-Based Approaches. *J. Afr. Earth Sci.* **2020**, *166*, 103826. [\[CrossRef\]](#)
24. Feng, R. A Bayesian Approach in Machine Learning for Lithofacies Classification and Its Uncertainty Analysis. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 18–22. [\[CrossRef\]](#)

25. Lin, J.; Li, H.; Liu, N.; Gao, J.; Li, Z. Automatic Lithology Identification by Applying LSTM to Logging Data: A Case Study in X Tight Rock Reservoirs. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1361–1365. [[CrossRef](#)]
26. Xie, D.; Liu, Z.; Wang, F.; Song, Z. A Transformer and LSTM-Based Approach for Blind Well Lithology Prediction. *Symmetry* **2024**, *16*, 616. [[CrossRef](#)]
27. Li, K.; Xi, Y.; Su, Z.; Zhu, J.; Wang, B. Research on Reservoir Lithology Prediction Method Based on Convolutional Recurrent Neural Network. *Comput. Electr. Eng.* **2021**, *95*, 107404. [[CrossRef](#)]
28. Wang, Z.; Zhao, R.; Yang, L.; Yin, H.; Tang, W.; Liu, D.; Gu, Y.; Jiang, Y. Paleo-Environmental and Geological Characteristics of Wufeng-Longmaxi Marine Shale in Different Paleo-Geomorphological Units, Eastern Sichuan Basin, China. *Nat. Gas Ind. B* **2025**, *12*, 572–584. [[CrossRef](#)]
29. Xiong, M.; Chen, L.; Gu, Z.; Chen, X.; Liu, B.; Lu, C.; Zhang, Z.; Wang, G. Pore Heterogeneity and Evolution of the Lower Silurian Longmaxi Shale Reservoir in the Southern Sichuan Basin: Responses to Sedimentary Environment. *Nat. Gas Ind. B* **2024**, *11*, 525–542. [[CrossRef](#)]
30. Yuan, C.; Zhao, D.; Heidari, A.A.; Liu, L.; Chen, Y.; Chen, H. Polar Lights Optimizer: Algorithm and Applications in Image Segmentation and Feature Selection. *Neurocomputing* **2024**, *607*, 128427. [[CrossRef](#)]
31. Tao, L.; Wang, Q.; Zhu, Z.; Yu, F.; Yin, X. NCLWO: Newton’s Cooling Law-Based Weighted Oversampling Algorithm for Imbalanced Datasets with Feature Noise. *Neurocomputing* **2024**, *610*, 128538. [[CrossRef](#)]
32. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [[CrossRef](#)]

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.