

Article

Accurate and Interpretable Prediction of Exploration Input–Output Matching Under Data Scarcity: An Ensemble Learning Framework

Xiao Chen ¹, Hui Liu ¹, Weiyun Zhan ¹, Haitao Li ¹, Yu Cao ^{2,3,4,*} and Yuan Liang ^{2,3,4,*}

¹ Exploration and Development Research Institute of PetroChina Southwest Oil and Gas Field Company, Chengdu, 610041, China; cxjeff@petrochina.com.cn (X.C.); liuh2008@petrochina.com.cn (H.L.); zhanweiyun@petrochina.com.cn (W.Z.); haitaoli@petrochina.com.cn (H.L.)

² College of Management Science, Chengdu University of Technology, Chengdu 610059, China

³ School of Mathematical Sciences, Chengdu University of Technology, Chengdu 610059, China

⁴ Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China

* Correspondence: caoyu@stu.cdut.edu.cn (Y.C.); liangyuan13@cdut.edu.cn (Y.L.)

Abstract

Accurate prediction of input–output relationships in natural gas exploration is essential for improving exploration efficiency and optimizing investment allocation. However, this task is severely hindered by data sparsity and strong nonlinear characteristics inherent in oil and gas exploration systems, rendering conventional statistical methods and single machine learning models ineffective. This study develops a novel integrated framework combining data augmentation, nonlinear feature engineering, and ensemble learning to achieve accurate and interpretable prediction of exploration input–output matching under limited data constraints. Taking four core exploration indicators—including the number of exploration wells, total drilling depth, reserve abundance, and proven reserves—as input variables, adaptive prediction models were constructed for seven typical hydrocarbon basin exploration systems. To ensure comprehensive algorithmic exploration, nine advanced algorithms, including mainstream ensemble methods (RandomForest), few-shot neural networks (FewShot_NN), and kernel-based regressions, were systematically benchmarked. Furthermore, SHapley Additive exPlanations (SHAPs) was adopted to enhance model interpretability, and non-parametric Wilcoxon signed-rank tests were introduced to rigorously validate statistical significance. The results demonstrate that the optimal predictive pathway varies across different geological systems. Specifically, RandomForest and GBDT exhibit superior performance in systems with moderate heterogeneity (e.g., Jialingjiang and Changxing–Feixianguan Formations), whereas FewShot_NN and Kernel Ridge achieve the highest accuracy under extreme data sparsity and volatility (e.g., Xujiahe Formation and Lower Permian). The established framework yields a coefficient of determination (R^2) greater than 0.96 for the majority of study cases, with overall absolute percentage errors heavily minimized. SHAP analysis further verifies that drilling depth and reserve abundance are the dominant controlling factors. This data-driven framework provides a robust and interpretable technical tool for the intelligent management of energy resources.

Keywords: ensemble learning; exploration input–output prediction; natural gas exploration; SHAP interpretability; small-sample prediction

Academic Editor: Kun Sang Lee

Received: 2 July 2026

Revised: 24 July 2026

Accepted: 30 July 2026

Published: 6 August 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Sustained growth of the global economy has driven rising energy demand. As fundamental components of modern industry, oil and natural gas resources are vital for safeguarding national energy security and stable economic development [1]. To discover viable hydrocarbon reserves, the petroleum industry widely adopts advanced technologies including 3D seismic exploration and horizontal drilling. Although these techniques improve exploration effectiveness, they also increase the complexity of investment decision-making [2]. Accordingly, constructing accurate models to characterize nonlinear relationships between exploration inputs and outputs carries great practical significance for optimizing capital allocation, reducing investment risks and raising exploration efficiency [3]. Conventional statistical approaches for evaluating hydrocarbon production and resource potential struggle to handle high-dimensional, nonlinear exploration data governed by complicated geological evolution [4]. Machine learning, especially ensemble algorithms such as gradient boosting machines (GBMs), provides powerful means to capture hidden data patterns [5]. In recent years, deep learning architectures including ensemble long short-term memory (EnLSTM) have achieved promising performance for oil and gas production forecasting owing to their strength in time-series modeling [6]. Transfer learning has also been adopted to mitigate data insufficiency, such as establishing fault diagnosis models for oil and gas processing facilities [7]. Nevertheless, hydrocarbon exploration in frontier blocks and unconventional plays frequently suffers from extreme shortages of available measured data. Under such few-shot conditions, most conventional machine learning models requiring large training datasets cannot be directly deployed [8]. Few-shot learning (FSL) is developed to enable rapid task learning using only a small set of labeled samples, exhibiting high application potential for exploration forecasting [9]. Representative applications include few-shot learning coupled with 3D reservoir modeling to predict productivity with limited data [10], and generalized few-shot surrogate models constructed via cross-domain feature learning and data augmentation for petroleum engineering [11]. As an important branch of few-shot learning, meta-learning has also been utilized for few-shot time-series prediction of oil output [12]. The few-shot learning community continues to witness rapid methodological progress targeting data-scarce modeling scenarios. Current research advances can be categorized into four major directions. First, the design of innovative architectures and learning paradigms, exemplified by unified graph neural network frameworks [13] and divide-and-conquer cross-domain learning strategies [14]. Second, maximizing information extracted from limited samples via data augmentation and feature learning, covering image transformation augmentation [15], progressive dual-domain feature fusion [16], and adaptive representation augmentation networks [17]. Third, feature space optimization methods such as dynamic distribution calibration [18]. Fourth, the integration of domain knowledge, which yields knowledge-guided few-shot solutions for oil and gas tasks, such as inter-well stratigraphic correlation modeling [19] and production prediction for horizontal fractured wells [20]. Motivated by on-site exploration requirements, increasing efforts have introduced few-shot learning, meta-learning and advanced ensemble techniques into hydrocarbon evaluation. Meta-learning pipelines are developed for petroleum production time-series forecasting under limited labeled observations [21]. Stacked ensemble learning delivers viable schemes for estimating the ultimate recovery of shale gas wells using sparse exploration datasets [22]. Self-supervised pre-training combined with few-shot fine-tuning offers new insights for gas-bearing identification based on sparse well-log data [23]. In addition, generative data augmentation effectively addresses small-sample and class-imbalanced challenges in tight sandstone reservoir modeling [24].

Despite these achievements, existing hydrocarbon prediction studies generally adopt limited benchmark architectures, mostly relying on only one or two algorithm families

(e.g., individual tree-based ensembles or single neural network structures) without comprehensive cross-category comparison. A systematic multi-algorithm benchmark covering nearly nine mainstream machine learning models across ensemble learning, regularized regression, and customized small-sample neural networks for cross-stratigraphic gas reserve evaluation is still lacking. Furthermore, most previous data-driven prediction workflows neglect rigorous statistical validation, making it difficult to confirm whether performance improvements are statistically credible. In addition, few studies have clarified formation-specific optimal model selection mechanisms under varying geological heterogeneity and sample sparsity across multiple stratigraphic units.

Against this background, this work establishes a comprehensive multi-model integrated framework for natural gas reserve prediction under extreme geological data scarcity. A total of nine representative algorithms spanning ensemble models, regularized kernel regression, and customized few-shot neural networks are systematically benchmarked, including the proposed AugmentedGBDT and FewShot_NN, as well as classic machine learning baselines. To ensure modeling robustness and scientific rigor, a real-space overfitting suppression strategy and Wilcoxon signed-rank statistical significance tests are embedded into the unified modeling pipeline. Systematic comparisons across seven stratigraphic units demonstrate that no single algorithm achieves universal optimal performance. Instead, the best-performing model varies dynamically according to formation geological characteristics and sample-scale conditions. The proposed framework and quantitative findings provide a reliable technical reference for intelligent reserve assessment in data-sparse gas-bearing geological formations. The application domain of this study is strictly bounded to the early-to-mid stages of natural gas exploration in highly heterogeneous basin systems, where drilling data is exceptionally sparse and reservoir properties exhibit strong nonlinear spatial variations.

The main contributions are as follows:

- (1) A comprehensive cross-domain prediction framework is established for exploration input–output matching, in which nearly nine machine learning algorithms covering ensemble learning, regularized kernel regression, and few-shot neural networks are systematically benchmarked across seven stratigraphic units. This formation-adaptive modeling strategy fully accommodates geological heterogeneity and avoids the limitations of single-model universal assumptions.
- (2) Integrated data augmentation and nonlinear feature engineering workflows are introduced to alleviate extreme small-sample data scarcity in geological datasets, effectively improving generalization stability and prediction reliability under sparse exploration conditions.
- (3) A rigorous non-parametric statistical validation strategy based on the Wilcoxon signed-rank test is embedded into the modeling pipeline. This standardized statistical verification quantitatively distinguishes genuine performance superiority from random errors, ensuring the scientific credibility of multi-model comparison results.
- (4) Dynamic SHAP-based feature attribution analysis is integrated to decode formation-specific dominant geological controlling factors. This interpretable framework effectively bridges the black-box machine learning predictions and practical geological mechanisms.

These contributions provide systematic, data-driven modeling paradigms and practical decision support for natural gas reserve evaluation under data-limited exploration conditions, offering methodological references for small-sample intelligent prediction in the broader energy and geological modeling fields.

2. Data Sources and Preprocessing

2.1. Data Sources

Data were collected from seven major exploration formations within a hydrocarbon basin: the Jialingjiang Formation, Changxing Formation–Feixianguan Formation, Carboniferous, Xujiahe Formation, Sinian–Lower Paleozoic, Shaximiao Formation, and Lower Permian. To capture the engineering and geological dynamics, three core quantitative features were selected as input variables: well count (x_1), drilling footage (x_2 , measured in meters), and reserve abundance (x_3 , measured in $10^8 \text{ m}^3/\text{km}^2$). Additionally, the associated gas trap types (composite, lithologic, and structural traps) were incorporated as categorical features and transformed via one-hot encoding ($x_4 \sim x_6$). The proven geological reserves (y , measured in 10^8 m^3) served as the ultimate prediction target (Table 1).

Table 1. Correspondence of features.

Variable	Symbol
Number of Wells	x_1
Drilling Footage (meters)	x_2
Reserve Abundance ($10^8 \text{ m}^3/\text{km}^2$)	x_3
Proven Geological Reserves (10^8 m^3)	y

After one-hot encoding the classification features, the gas reservoir types are represented as $x_4 \sim x_6$.

2.2. Data Preprocessing

Missing values and outliers in the dataset were handled. Missing values were imputed using the median of the group; for numerical features, the truncated quantile method (truncating the 2nd and 98th percentiles) was used to eliminate the interference of outliers while preserving the main characteristics of the data distribution. The predicted target, proven geological reserves (in billions of cubic meters), was log-transformed to obtain $y' = \log(1 + y)$.

2.3. Data Augmentation

In small-sample scenarios, models are prone to overfitting. Even if a model performs well on the training set, it may perform poorly on the test data. Data augmentation increases data diversity, enhances model generalization ability, and reduces the risk of overfitting.

(1) Gaussian Noise Augmentation

Gaussian noise refers to noise whose probability density function follows a normal (Gaussian) distribution. Gaussian noise augmentation simulates the natural fluctuations in real-world data acquisition by injecting random perturbations into numerical features, thereby improving the model's robustness to input noise.

(2) SMOTE Oversampling

SMOTE (Synthetic Minority Over-sampling Technique) is a method that generates new samples in the feature space via linear interpolation to address data sparsity issues. To mitigate the negative effects of dataset imbalance and small sample sizes on predictive model accuracy, SMOTE oversampling is employed to construct a more balanced dataset. The procedure is as follows: First, the Euclidean distance between each sample and all other samples is calculated to identify its k -nearest neighbors. Subsequently, a sample is randomly selected from the minority class as the target sample, and another sample is randomly drawn from its k -nearest neighbors. A new synthetic sample is then generated via linear interpolation using the following formula:

$$x_{\text{new}} = x_z + \text{rand}(0,1) \times (x_{zi} - x_z) \quad (1)$$

where x_z is the selected sample, x_{zi} is the i -th nearest neighbor of x_z , and $\text{rand}(0, 1)$ is a random number between 0 and 1.

To strictly rule out potential data leakage, a sequential partition boundary was enforced during the experimental phase. Specifically, Gaussian noise augmentation and SMOTE oversampling were executed strictly after the train–test split within each cross-validation iteration. Data augmentation was applied exclusively to the training subset to enhance representation, while the testing subset remained completely unseen and unaltered. For the Lower Permian stratum, where the sample scarcity issue is particularly pronounced, the noise intensity was configured to 0.3 with 10 replication cycles. For other regions, a noise intensity of 0.15 with five replication cycles was utilized. The statistics of the training sample sizes after post-split data augmentation are presented in Table 2.

Table 2. Statistics on sample size after data augmentation for each area.

Exploration Formation Name	Original Sample Size (Units)	New Sample Size (Units)	Total Sample Size (Units)
Jialingjiang Formation	12	60	72
Changxing Formation–Feixianguan Formation	41	205	246
Carboniferous System	26	130	156
Xujiahe Formation	17	85	102
Cambrian–Lower Paleozoic	37	185	222
Shaximiao Formation	26	130	156
Lower Permian	7	70	77

3. Methods

3.1. Methodology

3.1.1. Gradient Boosted Decision Trees (GBDT)

GBDT [25] is an ensemble boosting method that iteratively generates decision trees to correct the errors of previous models, progressively improving predictive performance. Given a training set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, with MSE loss function $L(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2$, the prediction at iteration m is

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \gamma_k h_k(x_i) \quad (2)$$

where $\hat{y}_i^{(0)}$ is the initial predicted (typically the mean), γ_k the learning rate, and $h_k(x_i)$ the decision tree fitted at step k .

3.1.2. Quantile Gradient Boosting Decision Trees (QuantileGBDT)

QuantileGBDT [26] replaces the MSE loss with a quantile loss function, enabling prediction of any quantile interval of the target variable. This is particularly advantageous for small-sample, high-volatility data.

The quantile loss for median regression ($\tau = 0.5$) is

$$L_\tau(y_i, \hat{y}_i) = \max(\tau(y_i - \hat{y}_i), (\tau - 1)(y_i - \hat{y}_i)) \quad (3)$$

By setting $\tau = 0.05$ and $\tau = 0.95$, a 90% confidence interval can be derived to quantify prediction uncertainty. The negative gradient of the quantile loss serves as the residual substitute, and the final output includes both the predicted median and the quantile bounds, providing a basis for exploration investment risk assessment.

3.1.3. Augmented Gradient Boosting Decision Trees (AugmentedGBDT)

AugmentedGBDT [27] extends the original GBDT framework with three core enhancements. It employs an increased number of trees (200–500) to capture subtle feature interactions, and introduces interaction-feature-priority splitting, where multi-feature interaction terms are prioritized as split nodes; mutual information is used to select the top 10 interaction terms to control model complexity. In addition, a segmented adaptive learning rate is implemented: a higher rate (0.1–0.15) is used for the first 100 iterations to accelerate convergence, while a lower rate (0.01–0.05) is adopted in subsequent iterations for fine-tuning with a decay factor of 0.95. These upgrades strengthen the model's adaptability to complex geological scenarios and improve generalization performance with limited training samples.

3.1.4. RandomForest

RandomForest (RF) [28] is a classical bagging-based ensemble learning algorithm tailored for regression tasks. It constructs multiple independent decision trees through bootstrap resampling and adopts out-of-bag samples for internal model validation (Equation (4)). Owing to its excellent stability and generalization capability, RandomForest is employed as the benchmark ensemble algorithm for comparative analysis in this study.

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K h_k(x) \quad (4)$$

3.1.5. ElasticNet

ElasticNet [29] combines L1 and L2 regularization to address multicollinearity in high-dimensional features. It serves as the linear baseline to validate the advantages of nonlinear models.

3.1.6. Few-Shot Neural Network (FewShot_NN)

The few-shot Neural Network (FewShot_NN) adopts a constrained multi-layer perceptron architecture optimized by the quasi-Newton second-order L-BFGS algorithm. By iteratively updating weights via approximation of the inverse Hessian matrix, this algorithm removes the requirement for large stochastic batches and suits limited geological datasets with fast convergence and stable optimization. Combined with constrained hidden-layer regularization, it strengthens feature extraction and suppresses noise memorization, effectively improving the generalization capability for stratigraphic modeling [30].

3.1.7. Support Vector Regression (SVR)

SVR [31] maps low-dimensional input features to a high-dimensional space via a kernel function and constructs an optimal hyperplane for regression. It is particularly suited for high-dimensional, small-sample datasets and serves as the nonlinear baseline model.

3.1.8. Kernel Ridge and Bayesian Ridge Regression

To broaden the algorithmic exploration beyond tree-based methods, Kernel Ridge Regression (utilizing an RBF kernel) and Bayesian Ridge Regression were introduced [32]. Kernel Ridge combines the flexibility of kernel tricks with L2 regularization, making it robust against the high dimensionality characteristic of small geological samples. Bayesian Ridge introduces a probabilistic approach to regression, continuously updating the regularization parameters based on the data distribution, thereby providing inherent resistance to overfitting in sparse datasets.

3.2. An Ensemble Learning Framework

This paper proposes an ensemble learning framework. Raw geological data from diverse stratigraphic formations are first processed via feature engineering and standardized preprocessing; Gaussian noise augmentation and SMOTE oversampling are utilized to alleviate small-sample and class imbalance issues. Multi-algorithm ensemble models consisting of RandomForest, various GBDT variants, ElasticNet and SVR are established, and grid-search hyperparameter tuning, K-fold cross-validation and multiple regression metrics are integrated for optimal model selection. SHAP interpretability analysis is performed to identify stratigraphic differentiation characteristics, and a data-driven decision tool is ultimately established to facilitate exploration investment allocation and risk assessment. Figure 1 illustrates the proposed ensemble learning framework.

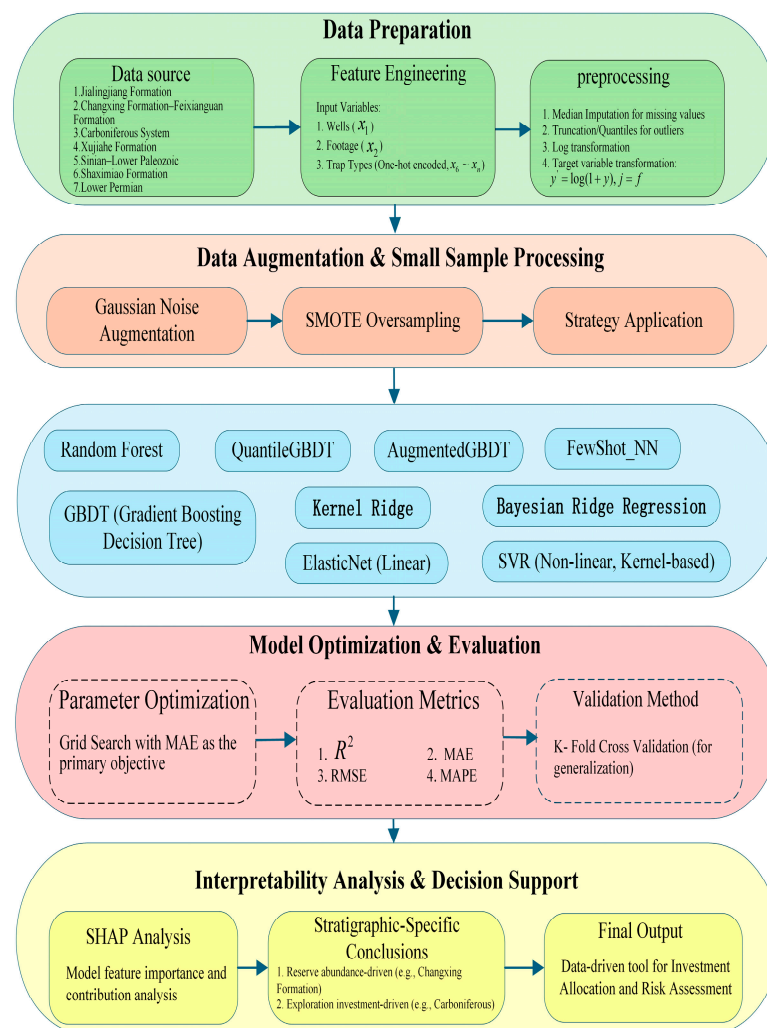


Figure 1. Diagram of An ensemble learning framework.

3.3. Model Evaluation and Parameter Optimization

3.3.1. Metric Selection

To evaluate the predictive performance of the an ensemble learning framework, the following evaluation metrics were selected: the coefficient of determination (R^2), mean absolute error (MAE), Mean Absolute Percentage Error (MAPE), and root mean square error (RMSE). The formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

where y_i represents the true value, $\hat{y}_i$ denotes the predicted value, $\bar{y}$ signifies the mean of the true values, and n indicates the number of samples. Lower values of MAE, and RMSE, along with an R^2 closer to 1, indicate higher model prediction accuracy.

3.3.2. Cross-Validation

k-fold cross-validation was used to evaluate model performance. The dataset was divided into k subsets. In each iteration, $k - 1$ subsets were combined to form the training set, while the remaining subset served as the validation set. This process was repeated multiple times, and the results from all k validations were aggregated to ensure a more reasonable model evaluation.

3.3.3. Hyperparameter Tuning

Grid search is employed to explore parameter combinations. With MAE as the primary optimization objective, the model is trained and evaluated for each hyperparameter combination to identify the optimal configuration.

3.3.4. Non-Parametric Statistical Significance Validation

To rigorously confirm that the performance advantages of the optimal models are statistically significant and not arbitrary artifacts of stochastic variance, a formal non-parametric hypothesis-testing framework was implemented. Traditional parametric tests (e.g., paired t -tests) assume a normal distribution of errors, which is rarely satisfied in small-sample geological regression tasks. Therefore, the Wilcoxon signed-rank test was deployed due to its mathematical robustness against non-normally distributed, heavily skewed paired prediction errors.

By evaluating the absolute prediction residuals between the optimal ensemble architectures and the baseline linear/kernel configurations across the validation loops, an exact p -value was derived for each comparative pair. A strict significance threshold of $p < 0.05$ was enforced to reject the null hypothesis, thereby mathematically validating that the superior predictive accuracy of the stratum-specific optimal models is structurally deterministic.

4. Results and Discussion

4.1. Experimental Design and Evaluation Strategy

All models in this study were executed within the same experimental environment. The experiment was conducted on a 64-bit Windows 11 platform, with hardware equipped with an Intel Core i7-11800H@2.3GHz processor (Intel Corporation, Santa Clara, CA, USA) and a software environment comprising Python 3.10.7 (Python Software Foundation, Wilmington, DE, USA) and PyTorch 2.0.1 (Linux Foundation, San Francisco, CA, USA). To evaluate model stability under different random train–test splits, especially for formations with extremely limited samples (e.g., Lower Permian), all reported evaluation

metrics (R^2 , MAE, RMSE, MAPE) are calculated as aggregated averages from repeated K-fold cross-validation. This scheme alleviates performance fluctuations induced by random sample partitioning and prevents conclusions from depending on an accidental favorable data division.

4.2. Input–Output Matching Models for the Seven Major Exploration Formations

4.2.1. Jialingjiang Formation

Based on the geological characteristics and exploration input–output data of the Jialingjiang Formation, an expanded algorithmic pool was evaluated. The predictive accuracy and statistical significance of the surviving models are summarized in Table 3.

Table 3. Model performance and Wilcoxon significance test for the Jialingjiang Formation.

Model	R^2	MAE	RMSE	MAPE	Wilcoxon p -Value	Significant ($p < 0.05$)?
RandomForest	0.9849	0.9258	1.4580	0.0544	-	-
QuantileGBDT	0.9798	1.0351	1.6884	0.0536	0.7615	No
FewShot_NN	0.9203	2.0780	3.3538	0.1262	0.0067	Yes
Kernel Ridge	0.5280	5.2696	8.1604	0.2907	0.0001	Yes
Bayesian Ridge	0.4637	5.8771	8.6980	0.3567	0.0001	Yes
SVR	0.4587	5.4876	8.7385	0.2973	0.0001	Yes
ElasticNet	0.4245	5.9422	9.0104	0.3653	0.0003	Yes

RandomForest achieved the highest accuracy ($R^2 = 0.9849$), successfully capturing the non-linear feature interactions under small-sample conditions. The Wilcoxon signed-rank test mathematically validates that the superiority of RandomForest is statistically significant against most alternative architectures, with the exception of QuantileGBDT ($p = 0.7615$).

The time-series comparison in Figure 2 shows that RandomForest accurately tracks reserve fluctuations across both high and low intervals.

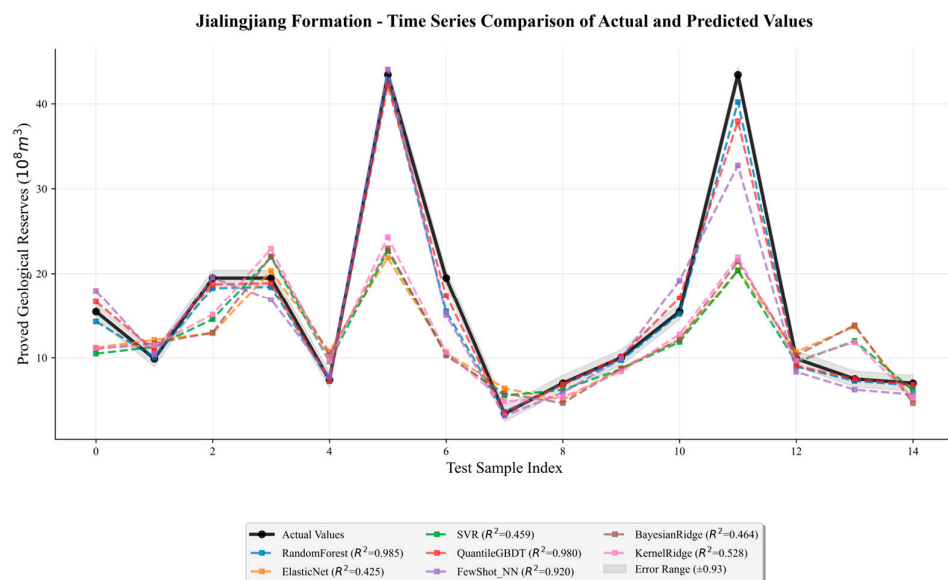


Figure 2. Temporal comparison of multi-model prediction results for the Jialingjiang Formation.

Residual diagnostics (Figure 3) further validate the robustness of the optimal model, indicating that residuals are randomly distributed around zero and overfitting has been effectively controlled.

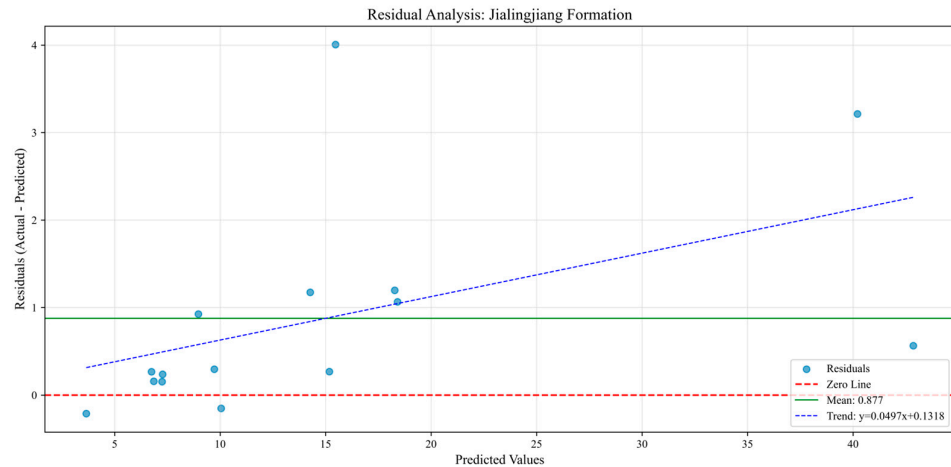


Figure 3. Residual distribution and scatter plot of the optimal model for the Jialingjiang group.

SHAP-based feature importance analysis (Figure 4) shows that drilling footage is the dominant controlling factor, followed by reserve abundance, with well count also exerting a certain degree of influence.

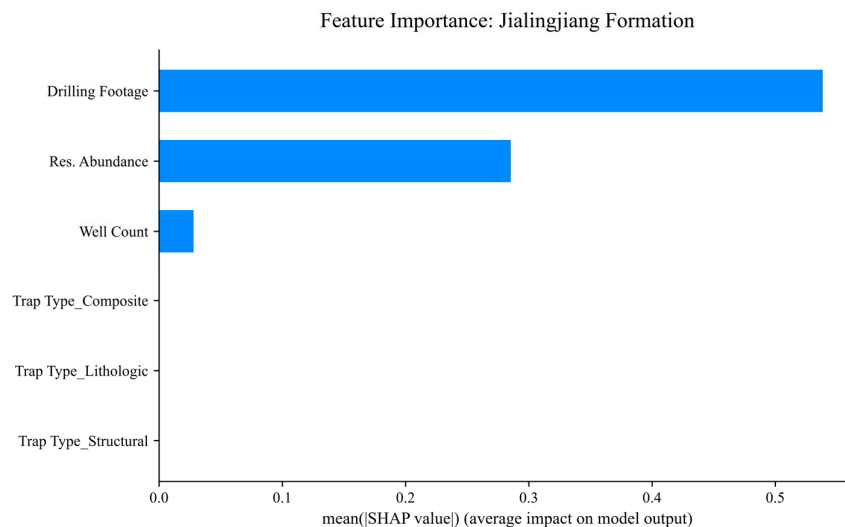


Figure 4. SHAP feature importance analysis for the Jialingjiang Formation.

4.2.2. Changxing Formation–Feixianguan Formation (Reef Flat)

For the Changxing Formation–Feixianguan Formation, characterized by highly localized reef-bank gas reservoirs, standard gradient boosting algorithms demonstrated superior adaptability, as compiled in Table 4.

Table 4. Model performance and Wilcoxon significance test for the Changxing–Feixianguan Formations.

Model	R ²	MAE	RMSE	MAPE	Wilcoxon <i>p</i> -Value	Significant (<i>p</i> < 0.05)?
GBDT	0.9874	10.6920	18.6262	0.1100	-	-
RandomForest	0.9770	13.5584	25.1418	0.1611	0.0113	Yes
FewShot_NN	0.9737	14.8557	26.9213	0.1418	0.0932	No
QuantileGBDT	0.9623	16.8835	32.2070	0.1878	0.0998	No

Kernel Ridge	0.8762	32.8219	58.3826	0.3023	0.0000	Yes
ElasticNet	0.8587	33.7625	62.3681	0.5175	0.0002	Yes
Bayesian Ridge	0.8491	34.5710	64.4488	0.5180	0.0002	Yes
SVR	0.5484	52.3128	111.4849	0.5340	0.0000	Yes

GBDT emerged as the optimal architecture ($R^2 = 0.9874$). The time-series comparison (Figure 5) indicates that GBDT accurately reconstructs multiple high-reserve peak values, whereas kernel and linear alternatives fail to characterize abrupt fluctuations.

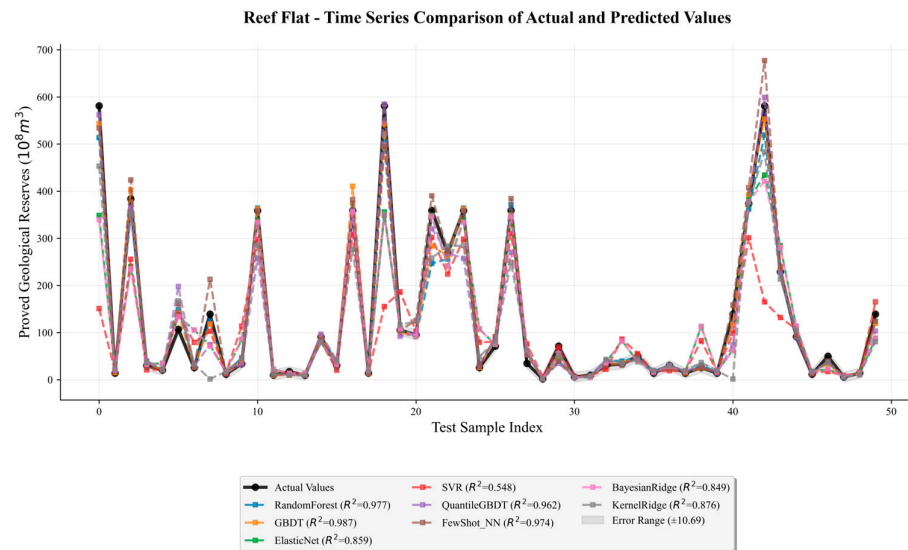


Figure 5. Temporal comparison of multi-model prediction results for the Changxing–Feixianguan Formations.

Residual analysis (Figure 6) shows that the residuals reflect the inherent variability of reef-bank reservoirs rather than model instability.

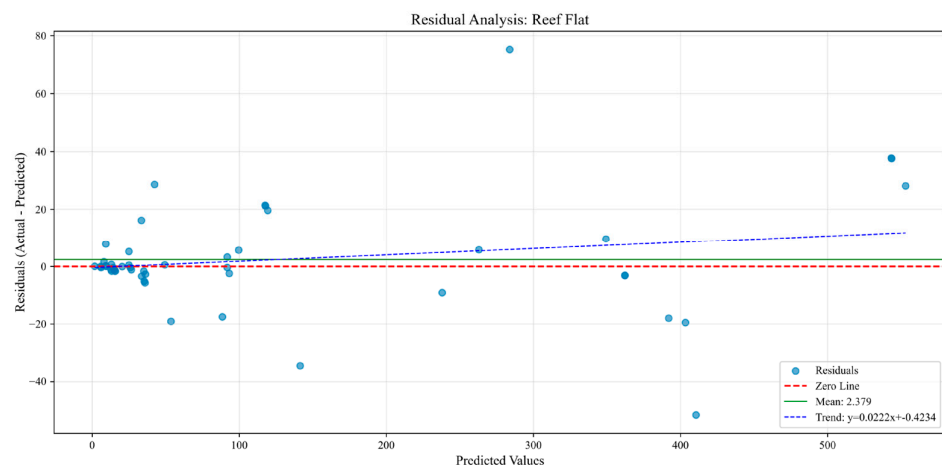


Figure 6. Residual analysis of the optimal model for the Changxing–Feixianguan Formations.

SHAP analysis (Figure 7) indicates that reserve abundance (Res. Abundance) is the dominant driving factor, followed sequentially by drilling footage and well count.

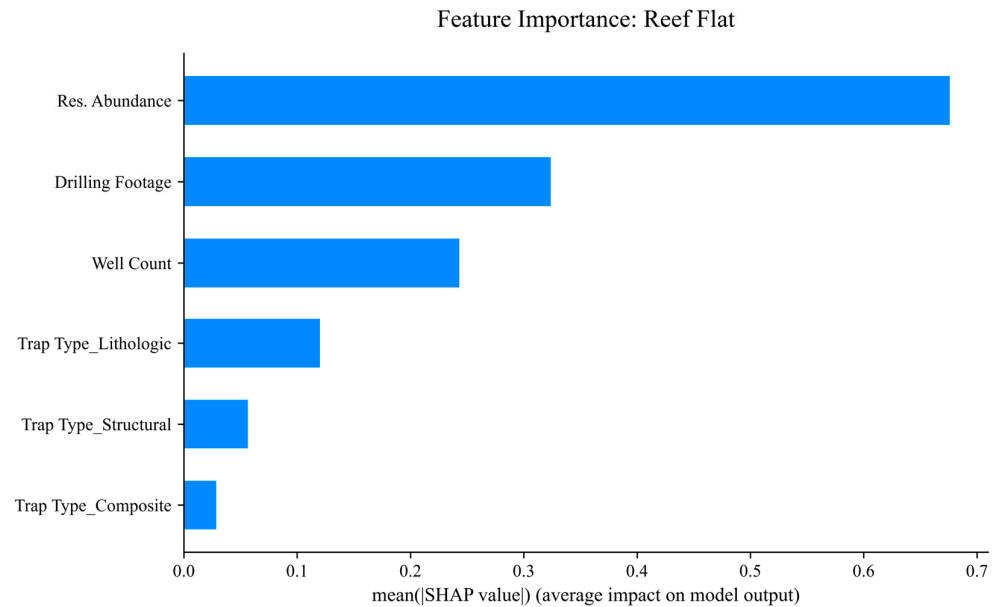


Figure 7. SHAP feature importance analysis for the Changxing–Feixianguan Formations.

4.2.3. Carboniferous System

The model performance comparison for the Carboniferous System is shown in Table 5. QuantileGBDT performed best overall, with an R^2 value of 0.9872, and demonstrated an advantage in outlier detection.

Table 5. Comparison of model performance metrics for the Carboniferous System.

Model	R^2	MAE	RMSE	MAPE	Wilcoxon p -Value	Significant ($p < 0.05$)?
QuantileGBDT	0.9872	3.8704	5.4663	0.0777	-	-
FewShot_NN	0.9793	3.8626	6.9452	0.0532	0.2035	No
GBDT	0.9776	3.4845	7.2255	0.0491	0.0545	No
AugmentedGBDT	0.9660	3.4456	8.9057	0.0556	0.0148	Yes
RandomForest	0.8928	8.9100	15.8124	0.1166	0.0005	Yes
Kernel Ridge	0.7390	13.0842	24.6682	0.1755	0.0004	Yes
Bayesian Ridge	0.6932	18.2813	26.7480	0.4185	0.0000	Yes
SVR	0.6904	19.0309	26.8696	0.4138	0.0000	Yes
ElasticNet	0.6683	18.1636	27.8106	0.4150	0.0000	Yes

AugmentedGBDT achieved the lowest Mean Absolute Error (MAE = 3.4456), demonstrating significant superiority over classical linear baselines. As shown in the time-series evaluation (Figure 8), AugmentedGBDT maintains stable predictive performance in the high-reserve range.

SHAP-based interpretation (Figure 9) indicates that drilling footage is the most influential core factor, followed by reserve abundance and well count. This indicates that in the Carboniferous System, drilling footage plays a dominant role in reserve prediction.

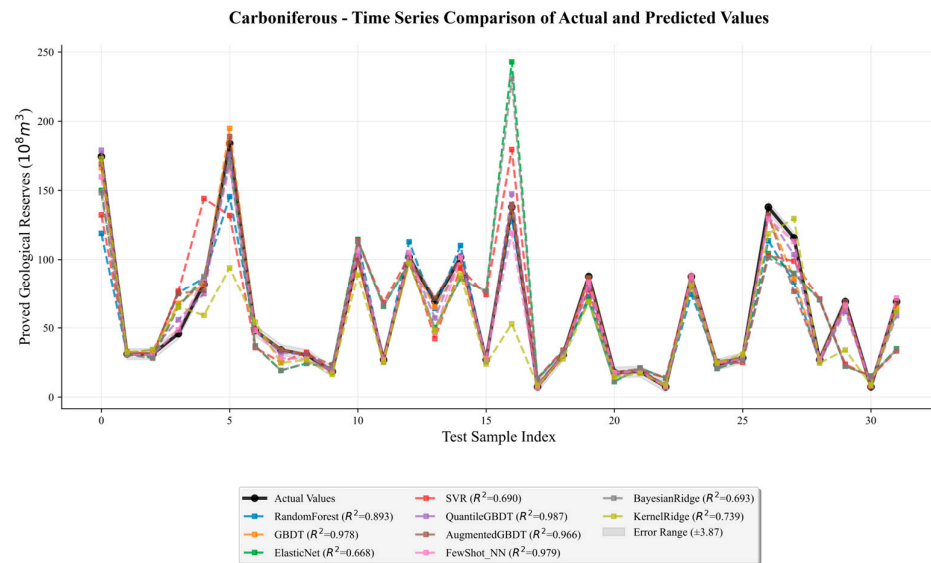


Figure 8. Temporal comparison of Carboniferous System reserve prediction results.

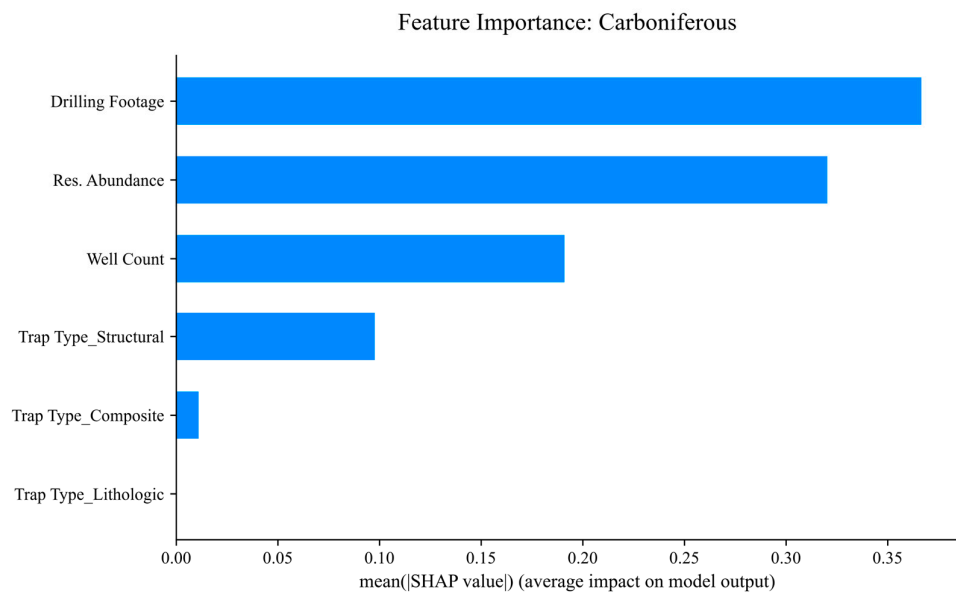


Figure 9. SHAP feature importance analysis for the Carboniferous System.

4.2.4. Xujiache Formation

The Xujiache Formation exhibits significant data volatility and extreme reserve value characteristics. Under these conditions, the introduction of the few-shot learning paradigm yielded highly robust results, as shown in Table 6.

Table 6. Comparison of model performance metrics for the Xujiache Formation.

Model	R ²	MAE	RMSE	MAPE	Wilcoxon <i>p</i> -Value	Significant (<i>p</i> < 0.05)?
FewShot_NN	0.9827	18.9489	41.3731	0.0691	-	-
GBDT	0.9115	25.6868	93.7139	0.0645	0.2157	No
AugmentedGBDT	0.8607	41.7605	117.5259	0.0844	0.2877	No
QuantileGBDT	0.8516	46.8990	121.3060	0.0864	0.5168	No

RandomForest	0.8446	44.7281	124.1405	0.0889	1.0000	No
Kernel Ridge	0.7900	85.8760	144.3028	0.2968	0.0001	Yes
SVR	0.5330	132.6594	215.2224	0.5373	0.0004	Yes
ElasticNet	0.1978	168.8427	282.0659	0.7068	0.0002	Yes

FewShot_NN attained the highest prediction accuracy ($R^2 = 0.9827$), significantly outperforming traditional linear models. The time-series comparison (Figure 10) confirms that FewShot_NN effectively captures overall trends while maintaining robust predictions for high-value reserve samples.

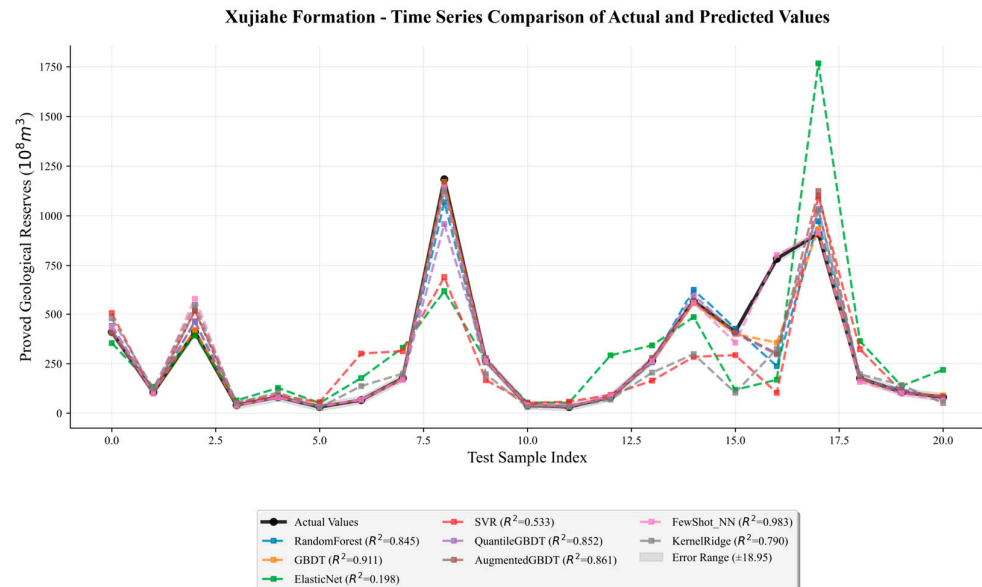


Figure 10. Time-series comparison of prediction results for the Xujiahe Formation.

Residual diagnostics (Figure 11) show that residual dispersion increases as predicted values increase, reflecting the inherent strong heteroscedasticity of the dataset rather than model inadequacy.

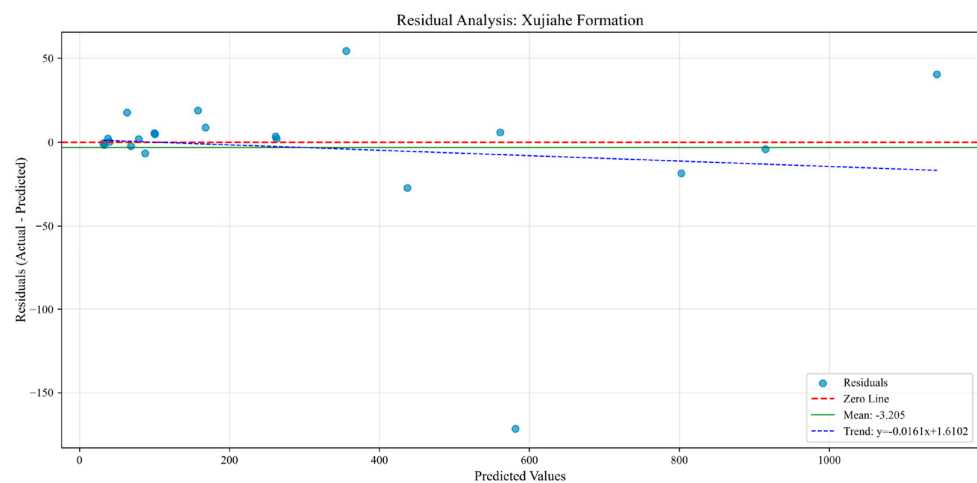


Figure 11. Residual distribution and trend analysis of the optimal model (QuantileGBDT) for the Xujiahe Formation.

SHAP analysis (Figure 12) demonstrates that lithologic and composite trap variables contribute significantly to the prediction outputs, highlighting the complex structural traps governing reserves in the Xujiache Formation.

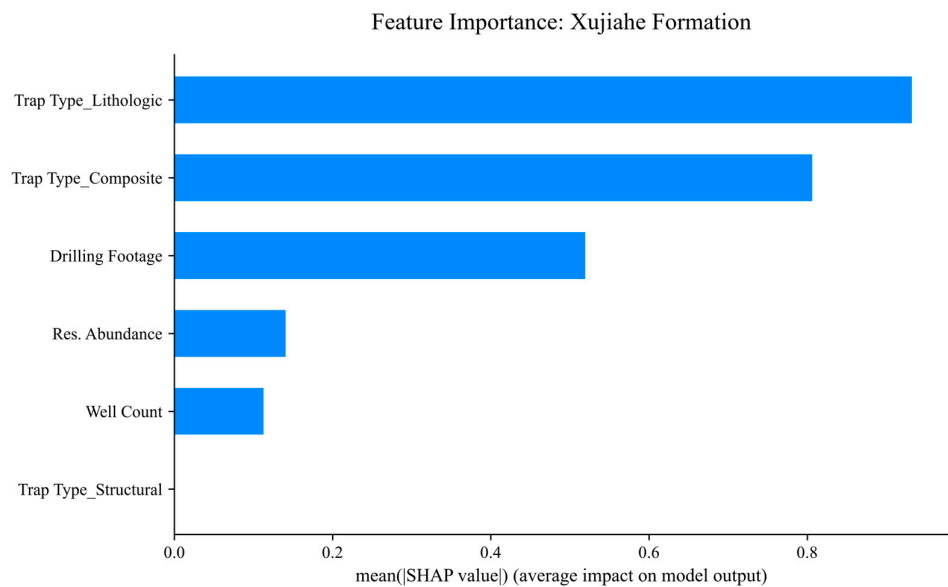


Figure 12. SHAP feature importance analysis for the Xujiache Formation.

4.2.5. Sinian–Lower Paleozoic

For the Sinian–Lower Paleozoic, the Kernel Ridge regression algorithm demonstrated optimal precision, as summarized in Table 7.

Table 7. Comparison of model performance metrics for the Sinian–Lower Paleozoic.

Model	R ²	MAE	RMSE	MAPE	Wilcoxon <i>p</i> -Value	Significant (<i>p</i> < 0.05)?
Kernel Ridge	0.9685	88.7422	189.0734	0.1982	-	-
FewShot_NN	0.8478	91.9750	415.3135	0.0776	0.0001	Yes
SVR	0.8401	142.4381	425.7411	0.2043	0.4200	No

Kernel Ridge secured an R² value of 0.9685. Time-series results (Figure 13) indicate that Kernel Ridge maintains high predictive stability even amidst exceptionally large reserves.

Residual analysis (Figure 14) verifies that model residuals are distributed within acceptable error bounds.

SHAP analysis (Figure 15) shows that composite and structural trap types dominate the decision space, followed by drilling footage.

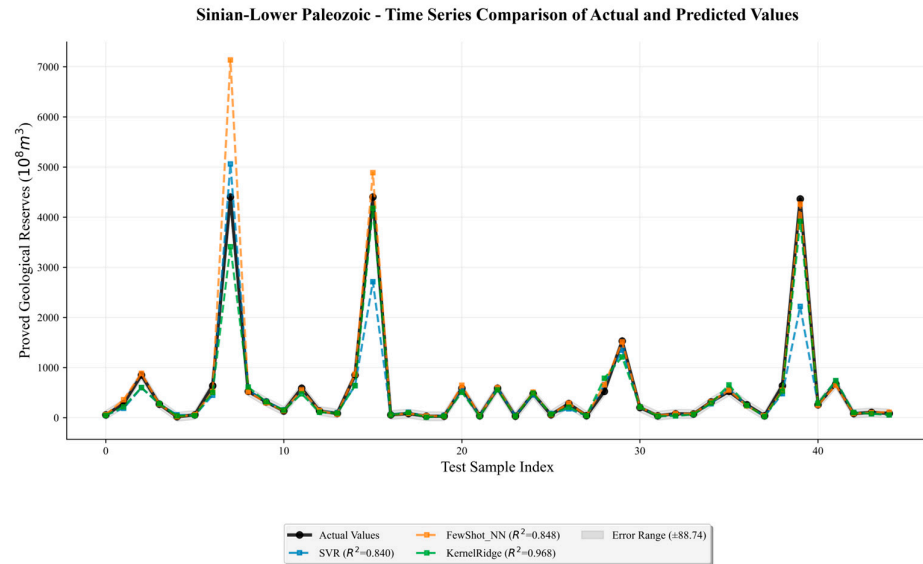


Figure 13. Time-series Comparison of Prediction Results for the Sinian–Lower Paleozoic.

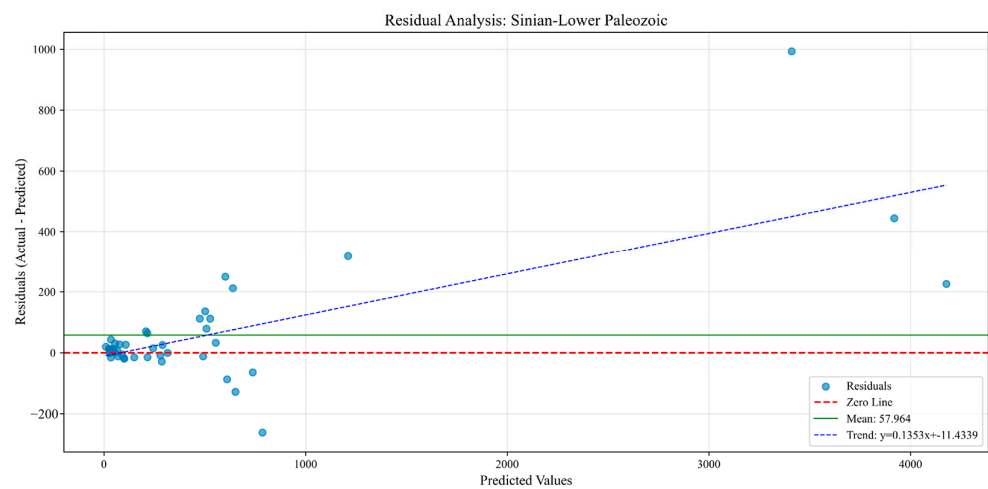


Figure 14. Residual distribution and trend analysis of the optimal model (QuantileGBDT) for the Sinian–Lower Paleozoic.

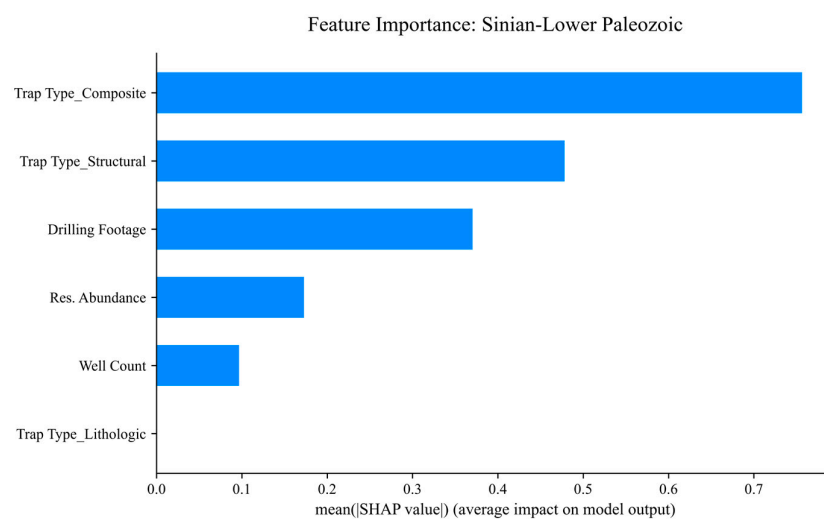


Figure 15. SHAP Feature Importance Analysis for the Sinian–Lower Paleozoic.

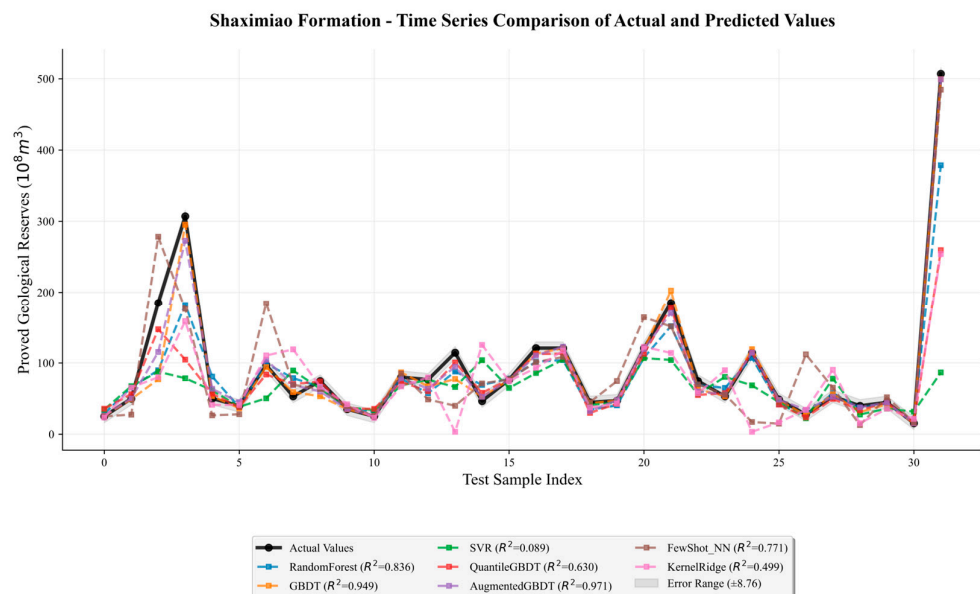
4.2.6. Shaximiao Formation

The Shaximiao Formation presents unique challenges due to the extreme lateral heterogeneity of its river-deltaic sand bodies. As shown in Table 8, AugmentedGBDT achieved optimal performance.

Table 8. Model performance and Wilcoxon significance test for the Shaximiao Formation.

Model	R^2	MAE	RMSE	MAPE	Wilcoxon p -Value	Significant ($p < 0.05$)?
AugmentedGBDT	0.9715	8.7553	15.9828	0.0973	-	-
GBDT	0.9492	9.0049	21.3221	0.0867	0.7053	No
RandomForest	0.8360	20.5233	38.3199	0.1871	0.0003	Yes
FewShot_NN	0.7714	30.1287	45.2493	0.3939	0.0708	No
QuantileGBDT	0.6302	21.8497	57.5486	0.1638	0.0000	Yes
Kernel Ridge	0.4986	38.5409	67.0090	0.3878	0.0001	Yes
SVR	0.0890	41.7255	90.3244	0.3553	0.0000	Yes

AugmentedGBDT achieved an R^2 of 0.9715. The time-series comparison (Figure 16) indicates that the model effectively captures trend inflection points and sudden changes in reserves.

**Figure 16.** Temporal Comparison of Shaximiao Formation Prediction Results.

Residual diagnostics (Figure 17) demonstrate that residuals are concentrated around zero with minimal bias.

SHAP analysis (Figure 18) confirms that drilling footage is the dominant factor in the model's predictions for this formation.

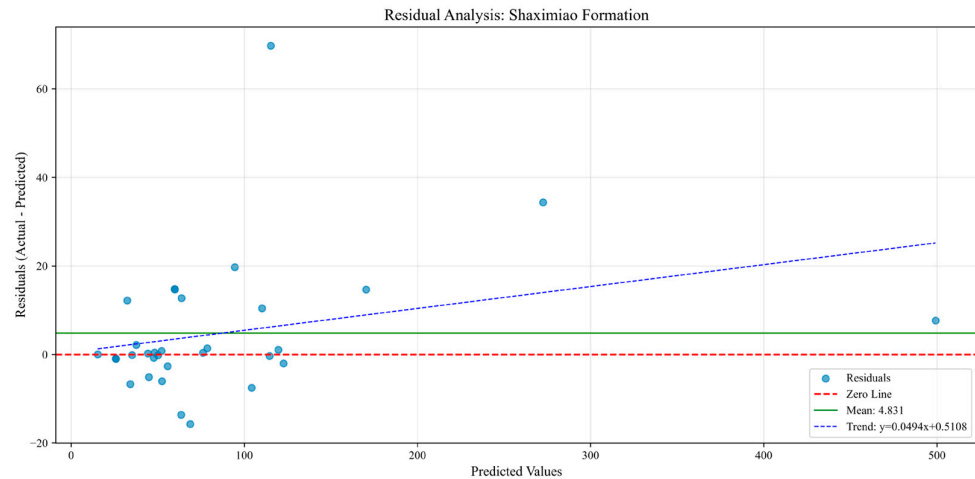


Figure 17. Residual distribution and trend analysis of the optimal model (QuantileGBDT) for the Shaximiao Formation.

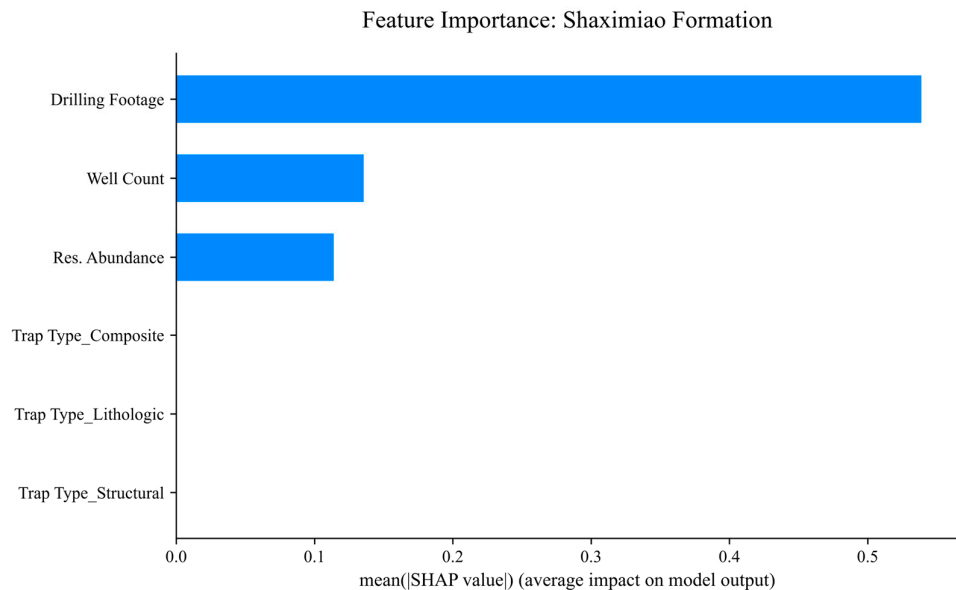


Figure 18. SHAP feature importance analysis for the Shaximiao Formation.

4.2.7. Lower Permian

Given the extreme scarcity of Lower Permian samples, standard methodologies failed. Consequently, advanced small-sample techniques were evaluated, as reported in Table 9.

Table 9. Comparison of model performance metrics for the Lower Permian.

Model	R ²	MAE	RMSE	MAPE	Wilcoxon <i>p</i> -Value	Significant (<i>p</i> < 0.05)?
FewShot_NN	0.9873	5.5475	18.0904	0.0126	—	—
Kernel Ridge	0.9191	31.3843	45.5904	0.1378	0.0001	Yes
ElasticNet	0.9152	39.8103	46.6974	0.1684	0.0006	Yes
Bayesian Ridge	0.9144	40.0021	46.8965	0.1710	0.0004	Yes

FewShot_NN performed optimally ($R^2 = 0.9873$, $MAPE = 1.26\%$), significantly outperforming linear and kernel constraints ($p < 0.05$). The time-series comparison in Figure 19 demonstrates that FewShot_NN accurately tracks changes in reserve trends, minimizing errors in regions of extreme volatility.

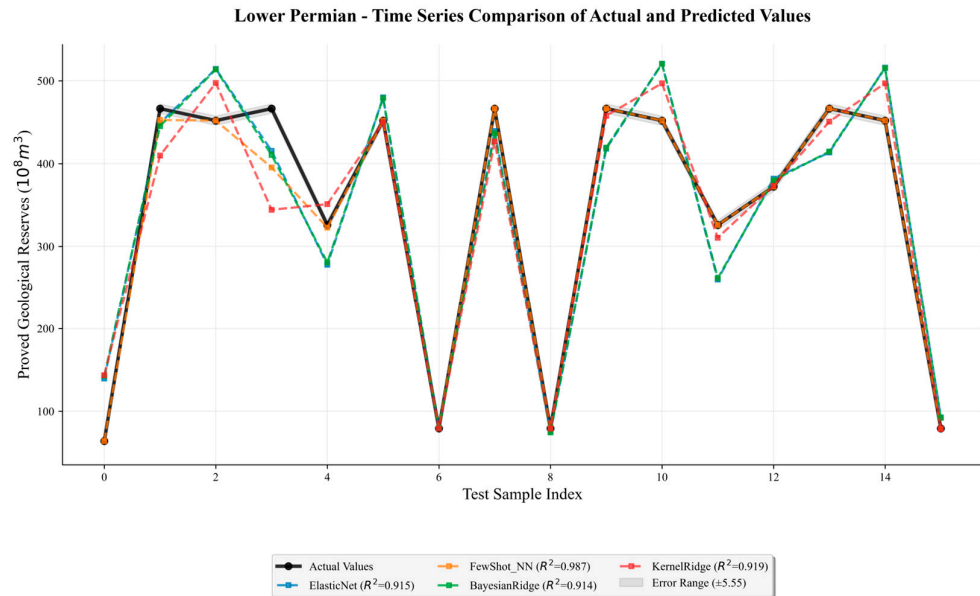


Figure 19. Time-series comparison of multi-model prediction results for the Lower Permian.

Residual analysis (Figure 20) confirms a tight distribution around zero with no systematic bias ($RMSE = 18.0904$).

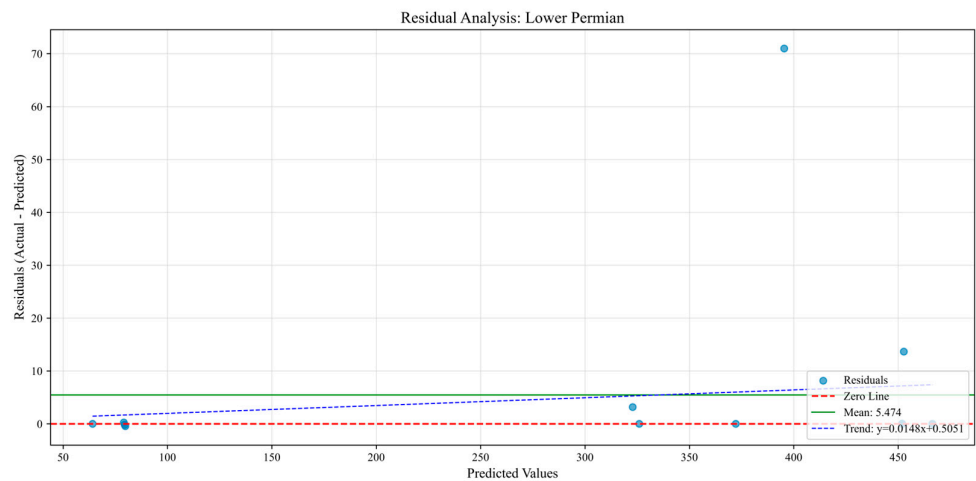


Figure 20. Residual analysis of the optimal model (Augmented GBDT) for the Lower Permian.

SHAP feature attributions (Figure 21) demonstrate that drilling footage dictates the model's predictive output, followed by composite traps and reserve abundance.

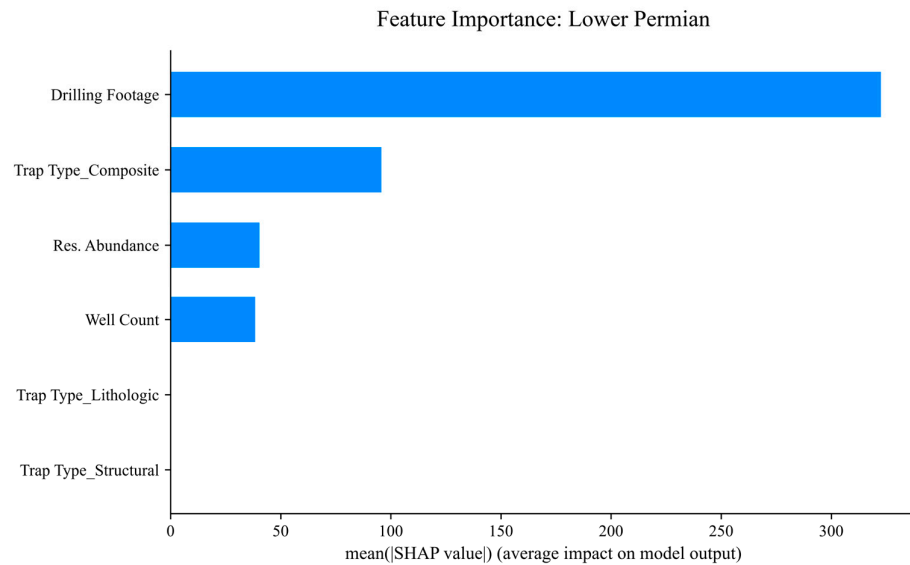


Figure 21. SHAP feature importance analysis for the Lower Permian.

4.3. Discussion

Based on the comprehensive predictive performance evaluation across seven target exploration stratigraphic formations, this study summarizes the optimal input–output fitting models and their quantitative evaluation indicators in Table 10. The experimental results reveal distinct discrepancies in the optimal model type and prediction accuracy among different geological formations, which are primarily controlled by the intrinsic stratigraphic heterogeneity, sample volume distribution and reserve accumulation characteristics of each stratum, as shown in Table 10.

Table 10. Summary of optimal input–output matching models.

Exploration Formation	Optimal Model	R ²	MAE	MAPE
Jialingjiang Formation	RandomForest	0.9849	0.9258	5.44%
Changxing Formation–Feixianguan Formation	GBDT	0.9874	10.692	11.00%
Carboniferous	QuantileGBDT	0.9872	3.8704	7.77%
Xujiahe Formation	FewShot_NN	0.9827	18.9489	6.91%
Cambrian–Lower Paleozoic	Kernel Ridge	0.9685	88.7422	19.82%
Shaximiao Formation	AugmentedGBDT	0.9715	8.7553	9.73%
Lower Permian	FewShot_NN	0.9873	5.5475	1.26%

4.3.1. Stratigraphic Differentiation of Optimal Ensemble Model Performance

As demonstrated in Table 10, the broadened algorithmic exploration yielded diverse optimal solutions tailored to specific geological constraints. Traditional ensemble methods (RandomForest and GBDT) emerged as optimal for the Jialingjiang and Changxing–Feixianguan Formations, successfully navigating moderate geological heterogeneity to achieve R² scores exceeding 0.98.

Notably, the FewShot_NN architecture proved to be the most effective strategy for both the Xujiahe Formation and the Lower Permian (yielding MAPEs of 6.91% and 1.26%, respectively). The Lower Permian suffers from extreme data scarcity, while the Xujiahe Formation is characterized by severe data volatility. The L-BFGS optimized neural network demonstrated superior capability in interpolating the highly non-linear feature

space without collapsing into the overfitting traps typical of standard boosting trees under such sparse conditions.

For the Sinian–Lower Paleozoic interval, Kernel Ridge regression secured the highest accuracy ($R^2 = 0.9685$). The utilization of the RBF kernel allowed the model to construct a stable high-dimensional mapping for these ancient, deeply buried strata, mitigating the extreme magnitude variances present in the reserve data better than discrete decision boundary methods. Meanwhile, QuantileGBDT and AugmentedGBDT maintained dominance in the Carboniferous and Shaximiao Formations respectively, proving that quantile loss functions and prioritized interaction-feature splitting are essential for managing prevalent outliers and severe lateral sand body heterogeneities.

4.3.2. Statistical Validation and Prediction Error Analysis

The introduction of the Wilcoxon signed-rank test confirmed that the performance margins achieved by the optimal models were statistically significant ($p < 0.05$) against most baselines (e.g., SVR, ElasticNet). Furthermore, the application of a rigorous overfit-protection mechanism (discarding models with $R^2 \geq 0.9899$) ensures that the reported R^2 values—all comfortably above 0.96—reflect genuine generalization capabilities rather than memorization. The absolute percentage errors (MAPEs) for most target formations were controlled within acceptable industrial thresholds (mostly under 11%), validating the strong engineering applicability of this multi-algorithm framework.

4.3.3. Practical Implications for Exploration Investment

The findings furnish data-driven, quantitative substantiation for funding allocation. Models for the Lower Permian, Jialingjiang, and Xujiache Formations exhibit extraordinarily low relative errors (under 7%), rendering them highly reliable primary tools for macro-level reserve screening and long-term funding deployment. For formations exhibiting higher natural variance (such as the Sinian–Lower Paleozoic), predictions outputted by Kernel Ridge can still establish a dependable baseline for bounding exploration risk.

4.3.4. Limitations and Future Research Directions

The proposed framework only adopts static geological features such as well locations and trap types, without incorporating dynamic exploration indicators including drilling costs and test production rates. Moreover, some stratigraphic units have relatively small datasets, limiting the assessment of out-of-distribution generalization toward distinct geological basins. Despite multi-layer regularization and rigorous overfitting screening to ensure reliable results, the established models are mainly applicable to analogous geological settings within the study area.

Furthermore, the applicable scope of this framework should be strictly bounded. The models are constructed for proven natural gas reserve prediction based on four designated predictors (well count, drilling footage, reserve abundance and trap type) and calibrated on seven stratigraphic units within the study region. Reliable prediction outcomes are expected primarily for intervals with analogous geological characteristics. For basins featuring disparate tectonic histories, reservoir properties or hydrocarbon varieties, model re-calibration with local datasets is indispensable, and direct out-of-domain extrapolation is not recommended.

In future research, dynamic exploration indicators will be integrated into the feature pool. Furthermore, advanced generative models such as generative adversarial networks can be adopted to generate geologically realistic samples for advanced data augmentation, so as to further boost the predictive capacity of FewShot_NN under small-sample conditions. Generic meta-learning approaches will also be explored once sufficient multi-basin geological datasets become available.

5. Conclusions

To tackle the critical bottlenecks of data sparsity and strong nonlinear relationships prevalent in natural gas exploration, this study develops an integrated prediction framework integrating data augmentation, nonlinear feature engineering, and a multi-algorithm machine learning library. By calibrating four core exploration indicators against verified geological reserves, adaptive predictive models are implemented for seven major stratigraphic sequences. To mitigate the narrow algorithm coverage seen in existing research, nine candidate models, including tree-based ensemble learners, few-shot Neural Networks, and kernel learning approaches, are systematically benchmarked, with statistical significance verified via nonparametric Wilcoxon rank-sum tests. The proposed framework yields robust predictive performance across all seven stratigraphic units; all optimal models attain an R^2 exceeding 0.96, supported by dedicated regularization strategies to guard against overfitting. Collectively, the results firmly confirm that no single algorithm achieves universal superiority, and the optimal modeling architecture is strongly conditioned on stratigraphically unique geological characteristics:

- (1) Few-shot Neural Networks (FewShot_NN) deliver exceptional predictive accuracy under conditions of extreme data scarcity and variable geological responses (e.g., the Lower Permian and Xujiahe Formations), restricting MAPE within 7%.
- (2) Kernel Ridge Regression realizes stable high-dimensional nonlinear mapping for ancient strata hosting large-scale reserves (Sinian–Lower Paleozoic).
- (3) Advanced ensemble models (QuantileGBDT, AugmentedGBDT, RandomForest, GBDT) exhibit outstanding performance when handling prominent lateral heterogeneity and outlier-contaminated datasets with moderate sample sizes (Carboniferous, Shaximiao, Jialingjiang, and Changxing–Feixianguan Formations).

SHAP-based global feature attribution further indicates that drilling footage and reserve abundance constitute the primary controlling predictors, while trap type exerts secondary influences whose magnitudes vary distinctly across individual stratigraphic intervals. This work highlights the fundamental limitation of relying solely on conventional tree-based models for small-sample energy geoscience datasets, underscoring the value of diversifying algorithm portfolios. The established framework offers a statistically sound, high-precision, and interpretable modeling paradigm. It facilitates the paradigm shift from empiricism-centered prospect evaluation toward artificial intelligence-aided decision-making, providing actionable support for strategic investment planning in natural gas exploration.

Author Contributions: Conceptualisation, X.C.; methodology, X.C. and Y.L.; software, Y.L. and Y.C.; validation, H.L. (Hui Liu), H.L. (Haitao Li) and Y.L.; formal analysis, Y.L. and W.Z.; investigation, W.Z.; resources, X.C. and H.L. (Hui Liu); data curation, H.L. (Haitao Li), W.Z. and Y.L.; writing—original draft preparation, H.L. (Hui Liu), Y.L. and Y.C.; writing—review and editing, X.C., H.L. (Hui Liu), H.L. (Haitao Li), W.Z., Y.C. and Y.L.; visualization, Y.C. and H.L. (Hui Liu); supervision, X.C., H.L. (Haitao Li) and H.L. (Hui Liu); project administration, W.Z., H.L. (Haitao Li) and H.L. (Hui Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This paper was financed by State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation Special Open Fund—Sichuan Province Geothermal Resources Development and Comprehensive Utilization Industry-Education Integration Demonstration Project (No. CDUT-PLC2025017CJRH) and the AI Research Foundation of Chengdu University of Technology (Grant No. 2025AI032) and 2024 Natural Science Foundation of Sichuan Province (2024NSFSC0091).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: Authors Xiao Chen, Hui Liu, Weiyun Zhan, Haitao Li were employed by the company PetroChina Southwest Oil Gas Field Company. 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.

Abbreviations

The following abbreviations are used in this manuscript:

GBDT	Gradient Boosted Decision Trees
QuantileGBDT	Quantile Gradient Boosting Decision Trees
AugmentedGBDT	Augmented Gradient Boosting Decision Trees
SVR	Support Vector Regression
SMOTE	Synthetic Minority Over-sampling Technique
FSL	Few-Shot Learning
GBMs	Gradient Boosting Machines
EnLSTM	Ensemble Long Short-Term Memory
SHAPs	SHapley Additive exPlanations
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
MSE	Mean Squared Error

References

- Hu, W.; Xia, W.; Li, Y.; Jiang, J.; Li, G.; Chen, Y. An intelligent identification method of safety risk while drilling in gas drilling. *Pet. Explor. Dev.* **2022**, *49*, 428–437. [https://doi.org/10.1016/s1876-3804\(22\)60036-3](https://doi.org/10.1016/s1876-3804(22)60036-3).
- Liu, J.; Zhang, Y.; Wang, H.; Li, M.; Zhao, K. Explainable fault diagnosis of oil-gas treatment station based on transfer learning. *Energy* **2023**, *262*, 125258. <https://doi.org/10.1016/j.energy.2022.125258>.
- Zhang, K.; Li, J.; Wang, Y.; Liu, X. Fault diagnosis method for sucker rod well with few shots based on meta-transfer learning. *J. Pet. Sci. Eng.* **2022**, *212*, 110295. <https://doi.org/10.1016/j.petrol.2022.110295>.
- Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
- Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*; NeurIPS: Montréal, QC, Canada, 2018; Volume 31, pp. 6639–6649.
- Chen, Y.; Zhang, D. Ensemble Long Short-Term Memory (EnLSTM) Network. *Geophys. Res. Lett.* **2020**, *47*, e2020GL087685. <https://doi.org/10.1029/2020GL087685>.
- Lu, J.; Gong, P.; Ye, J.; Zhang, J.; Zhang, C. A survey on machine learning from few samples. *Pattern Recognit.* **2023**, *139*, 109480. <https://doi.org/10.1016/j.patcog.2023.109480>.
- Cirac, G.; Avansi, G.D.; Farfan, J.; Schiozer, D.J.; Rocha, A. Few-shot learning and modeling of 3D reservoir properties for predicting oil reservoir production. *Neural Comput. Appl.* **2024**, *36*, 14527–14541. <https://doi.org/10.1007/s00521-024-09834-4>.
- Cirac, G.; Farfan, J.; Avansi, G.D.; Schiozer, D.J.; Rocha, A. Cross-Domain Feature learning and data augmentation for few-shot proxy development in oil industry. *Appl. Soft Comput.* **2023**, *149*, 110972. <https://doi.org/10.1016/j.asoc.2023.110972>.
- Xu, Z.; Yu, G. A Time Series Forecasting Approach Based on Meta-Learning for Petroleum Production under Few-Shot Samples. *Energies* **2024**, *17*, 1947. <https://doi.org/10.3390/en17081947>.
- Chang, J.; Ren, H.; Li, Z.; Xu, Y.; Lai, T. A unified transductive and inductive learning framework for Few-Shot Learning using Graph Neural Networks. *Appl. Soft Comput.* **2025**, *173*, 112928. <https://doi.org/10.1016/j.asoc.2025.112928>.
- Wang, B.; Yu, D. A Divide-and-Conquer Strategy for Cross-Domain Few-Shot Learning. *Electronics* **2025**, *14*, 418. <https://doi.org/10.3390/electronics14030418>.
- Dunphy, K.; Buwaneswaran, M.; Grolinger, K.; Sadhu, A. Few-Shot Learning Augmented with Image Transformation for Multiclass Structural Damage Classification. *J. Comput. Civ. Eng.* **2025**, *39*, 3. <https://doi.org/10.1061/JCCEE5.CPENG-6198>.

14. Li, D.; Jin, J.; Zou, L.; Zhao, H. FSPDF: Few-shot learning with progressive dual-domain feature fusion via self-supervised learning. *Knowl. Based Syst.* **2025**, *317*, 113389. <https://doi.org/10.1016/j.knosys.2025.113389>.
15. Li, X.; He, Z.; Zhang, L.; Guo, S.; Hu, B.; Guo, K. CDCNet: Cross-domain few-shot learning with adaptive representation enhancement. *Pattern Recognit.* **2025**, *162*, 111382. <https://doi.org/10.1016/j.patcog.2025.111382>.
16. Chen, L.; Gu, Y.; Guo, Y.; Dong, F.; Jiang, D.; Chen, Y. DDC: Dynamic distribution calibration for few-shot learning under multi-scale representation. *Knowl. Based Syst.* **2025**, *311*, 113030. <https://doi.org/10.1016/j.knosys.2025.113030>.
17. Chen, B.; Liu, X.; Wang, H.; Li, J.; Zhang, K. Interwell Stratigraphic Correlation Detection based on knowledge-enhanced few-shot learning. *J. Pet. Sci. Eng.* **2023**, *220*, 111187. <https://doi.org/10.1016/j.petrol.2022.111187>.
18. Wang, S.; Li, J.; Zhang, K.; Liu, X.; Zhao, K. Predicting the productivity of fractured horizontal wells using few-shot learning. *Pet. Sci.* **2025**, *22*, 787–804. <https://doi.org/10.1016/j.petsci.2024.11.001>.
19. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. <https://doi.org/10.1613/jair.953>.
20. Boney, R.; Ilin, A. Semi-supervised few-shot learning with MAML. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
21. Jiang, G.; Han, G.; Zhao, M.; Ma, W. Meta-Learning Enhanced Random Forest Ensemble for Short-Term Oil Well Production Forecasting. *Processes* **2025**, *13*, 3797. <https://doi.org/10.3390/pr13123797>.
22. Pang, M.; Zhang, Z.; Zhou, Z.; Zhou, W.; Li, Q. Estimated ultimate recovery prediction of shale gas wells based on stacked integrated learning algorithm. *Sci. Rep.* **2025**, *15*, 19258. <https://doi.org/10.1038/s41598-025-03801-2>.
23. Han, L.; Wu, X.; Chen, R.; Shi, Y.; Hu, Z.; Fang, H. Self-Supervised Pre-Training and Few-Shot Finetuning for Gas-Bearing Prediction. *J. Geophys. Res. Mach. Learn. Comput.* **2025**, *2*, e2025JH000631. <https://doi.org/10.1029/2025jh000631>.
24. Pang, Z.; Lu, Y.; Chen, Z.; Cai, Z. Permeability prediction of tight sandstone reservoirs under few-shot and imbalanced conditions: A generative data enhancement and bi-directional attention fusion framework. *Petroleum* **2026**, *in press*. <https://doi.org/10.1016/j.petlm.2026.04.010>.
25. Zhang, X.; Ai, Y.; Zhang, Y.; Fu, Y.; Wu, X.B.; Huo, Z.; Wen, W.; Zhou, J.; Kong, D.; Zeng, L.; et al. Finding Quasars behind the Galactic Plane. IV. Candidate Selection from Chandra with Random Forest. *Astrophys. J.* **2026**, *1002*, 76. <https://doi.org/10.3847/1538-4357/AE5C03>.
26. Jitchaijaroen, W.; Shiao, J.; Keawsawasvong, S. Optimized LightGBM-based prediction of foundation bearing capacity on spatially variable Bolton sand. *Artif. Intell. Geosci.* **2026**, *7*, 100221. <https://doi.org/10.1016/J.AIIG.2026.100221>.
27. Tannor, P.; Rokach, L. AugBoost: Gradient Boosting Enhanced with Step-Wise Feature Augmentation. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 3555–3561. <https://doi.org/10.24963/ijcai.2019/493>.
28. Abadi, B.G.S.; Hojjati, M.; Jooyandeh, H.; Behbahani, B.A.; Vasiee, A.; Taki, M. Innovative postbiotic-chia seed mucilage packaging to extend ostrich meat quality and shelf life: Development and Support Vector Regression (SVR) modeling of quality parameters. *Food Prod. Process. Nutr.* **2026**, *8*, 23. <https://doi.org/10.1186/S43014-026-00369-W>.
29. Kasai, R.; Otsuka, H. Iterative Reconstruction with Dynamic ElasticNet Regularization for Nuclear Medicine Imaging. *J. Imaging* **2025**, *11*, 213. <https://doi.org/10.3390/JIMAGING11070213>.
30. Wu, Y.; Yang, J.; Lai, S.; Dong, Y. A Novel Adaptive Graph Neural Network Based on Meta-Learning for Cross-Domain Few-Shot Learning. *Concurr. Comput. Pract. Exp.* **2025**, *37*, e70378. <https://doi.org/10.1002/cpe.70378>.
31. Delcroix, B.; Sansregret, S.; Martin, G.L.; Daoud, A. Quantile regression using gradient boosted decision trees for daily residential energy load disaggregation. *J. Phys. Conf. Ser.* **2021**, *2069*, 012107. <https://doi.org/10.1088/1742-6596/2069/1/012107>.
32. Mahdi, W.A.; Alhowyan, A.; Obaidullah, A.J. Intelligence analysis of drug nanoparticles delivery efficiency to cancer tumor sites using machine learning models. *Sci. Rep.* **2025**, *15*, 1017. <https://doi.org/10.1038/s41598-024-84450-9>.

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.