Petroleum Science 23 (2026) 7001–7016

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A KAN-based Adaptive Self-training Semi-supervised Ensemble Learning (KASSEL) framework for lithology identification in unconventional reservoirs

Yu-Xuan Deng [a,b], Wen-Dong Wang [a,b,*], Feng Jia [a,b], Yu-Liang Su [a,b], Xin-Yu Zhuang [a,b], Cun-Qi Jia [c], Yong-Mao Hao [a,b], Lei Li [a,b]

[a] State Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[b] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[c] Physical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia

ARTICLE INFO

Article history:
Received 15 December 2025
Received in revised form
5 March 2026
Accepted 5 September 2026
Available online 10 September 2026

Edited by Xiu-Fang Hu

Keywords:
Lithology identification
Kolmogorov-Arnold networks
Semi-supervised learning
Self-training algorithm

ABSTRACT

Lithology identification is a fundamental task in unconventional petroleum exploration, and acquiring accurate lithological information is vital for reservoir development. However, this task faces three interconnected challenges: limited labeled samples, severe class imbalance, and poor model interpretability, which collectively restrict classification accuracy and practical application. To address these challenges, this paper proposes a KAN-based Adaptive Self-training Semi-supervised Ensemble Learning (KASSEL) framework for lithology identification. The KASSEL framework leverages Kolmogorov-Arnold Networks (KAN) as base classifiers and introduces a novel two-stage ADASYN algorithm to handle class imbalance issues. By integrating adaptive sampling with self-training semi-supervised learning, the framework effectively utilizes both labeled and unlabeled data to build a robust ensemble system. Experimental results demonstrate that KASSEL achieves a macro F1-score of 0.89 on the test dataset, significantly outperforming conventional machine learning and deep learning methods. The framework substantially improves minority class recognition where conventional methods often fail, while maintaining enhanced interpretability through SHAP-based post-hoc analysis combined with KAN's learnable activation functions. The framework has been successfully applied to the S tight reservoir, achieving high-precision lithology classification and providing engineers with an efficient workflow for rapid lithology determination. The KASSEL framework offers an efficient, accurate, and practically viable solution for lithology identification in unconventional reservoir exploration and development.
© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

In engineering geology practice, lithology identification transcends basic stratigraphic categorization to become a critical control parameter for subsurface engineering operations. Accurate lithology identification constitutes a fundamental pillar in petroleum exploration and reservoir characterization across all hydrocarbon play types (Barach et al., 2021; Krivoshchekov et al., 2023; Mishra et al., 2022; Xie et al., 2018; Zhang et al., 2026). From conventional sandstone reservoirs to unconventional shale plays, precise determination of rock types governs key decisions in well placement, completion design, and production forecasting (Ali et al., 2025; Zhuang et al., 2025). Particularly in the era of energy transition, the ability to discern subtle lithological variations becomes paramount. The International Energy Agency emphasizes that advanced subsurface characterization technologies form the cornerstone of low-carbon drilling strategies, with lithology identification serving as a critical component in minimizing unnecessary well construction (IEA, 2021). This imperative holds heightened significance in unconventional reservoirs, which are influenced by various geological factors such as sedimentation, tectonic evolution, and diagenesis that result in complex, diverse, and heterogeneous lithology. This complexity creates a significant

* Corresponding author
E-mail address: wwdong@upc.edu.cn (W.-D. Wang).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.09.007
1995-8226/© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

challenge for lithology identification (Liu et al., 2023; Venieri et al., 2021; Zhang et al., 2022b). Conventional methods of lithology identification involve laboratory analysis of core samples and geologists' analysis of well logs (Fang et al., 2024; Karimi et al., 2021; Yin et al., 2018). Although core samples offer the most reliable lithology information, the prohibitively high cost and time-intensive nature of coring operations limit comprehensive lithological characterization through either core samples or cuttings logging (Dong et al., 2022; Grana et al., 2012; Ilkhchi et al., 2006; Ma, 2011). Meanwhile, geophysical log interpretation often shows significant inter-expert variability in complex lithofacies (Chang et al., 2000; Maiti and Tiwari, 2010). These limitations create critical knowledge gaps that propagate through reservoir models, particularly in undrilled intervals where real-time decisions rely heavily on empirical extrapolation (Lu et al., 2023). Although continuous measurements are provided by logging data, the mineralogical complexities exhibited by many unconventional reservoirs pose challenges to traditional interpretation frameworks initially developed for conventional plays.

Lithology identification is a nonlinear classification problem that requires nonlinear characteristic extraction and application. In recent years, machine learning methods have been extensively applied to lithology identification tasks considering their efficiency and accuracy (Han et al., 2023; Liu et al., 2020; Sun et al., 2020). These techniques have proven effective for lithology identification by establishing mathematical relationships between lithology and log data via rapid, repetitive computations. Studies have proved that both supervised and unsupervised learning methods (e.g., support vector machines, neural networks, random forests) achieve remarkable success in classifying various rock types (Bhattacharya et al., 2016; Gu et al., 2019; Leite et al., 2013; McCreery and Al-Mudhafar, 2017; Rosid et al., 2019; Saporetti et al., 2018; Zhu et al., 2018). However, despite these advances, current research faces several interconnected challenges that limit practical application. More recently, Dong et al. (2023a) proposed a deep kernel method that enhances kernel principal component analysis with residual network structures, achieving superior accuracy in complex lithofacies identification from conventional well logs compared to traditional kernel approaches. Dong et al. (2023b) further investigated practical ensemble strategies and principles for lithofacies identification, demonstrating that both homogeneous and heterogeneous ensemble approaches can effectively reduce prediction variance and bias errors, with heterogeneous ensemble being particularly suitable for constructing powerful identification models.

A major constraint in lithology identification lies in the high cost of core data acquisition, resulting in limited sample sizes for model training. The problem of limited samples substantially increases the chance of overfitting, particularly when attempting to define formation lithology in undrilled intervals. Recent work by Song et al. (2023) proposes an adversarial learning-based approach to mitigate overfitting from limited data, while Tolstaya et al. (2023) demonstrates how data cleaning techniques can enhance model generalization with sparse samples. In a complementary direction, Dong et al. (2024) introduced a semi-supervised ladder network that leverages both labeled and unlabeled well-log data for lithofacies identification, achieving over 6% accuracy improvement compared to purely supervised models when labeled samples are scarce, highlighting the potential of semi-supervised frameworks in geological applications with limited annotations. Notwithstanding these initiatives, the development of more effective small-sample learning frameworks is still an essential requirement for practical applications in geological settings where comprehensive data collection is usually prohibitively costly.

This problem of data shortage is further compounded by inherent class imbalance in geological formations. The natural dearth of specific lithologies in unconventional reservoirs leads to the creation of imbalanced multiclass scenarios that significantly degrade classification accuracy. With such skewed distributions, conventional classifiers suffer notable performance reduction. He et al. (2020) addresses this through MAHAKIL oversampling, significantly improving $F\beta$-scores in tight sandstone reservoirs. Meanwhile, Zhang et al. (2022a) employs cross-validation and Bayesian optimization to manage multiclass challenges, highlighting the critical role of sample balancing techniques. Nevertheless, more advanced algorithms are still needed to deal with extreme imbalance situations and high-dimensional multiclass issues related to numerous lithotypes, particularly when working with inherently scarce geological samples.

Building upon these technical challenges, perhaps the most critical barrier to widespread adoption remains the limited interpretability of machine learning models in geological contexts. The "black-box" nature of conventional machine learning approaches, particularly deep neural networks, significantly limits practical application. Lee and Lumley (2023) emphasize this need for transparency in their work on interpretable deep learning for shale brittleness prediction. Similarly, Al-Mudhafar (2017) demonstrates the advantages of more explainable architectures like probabilistic neural networks in lithology identification. This interpretability challenge becomes especially pronounced when models must perform reliably with small, imbalanced datasets.

To address these challenges, this paper presents KASSEL (KAN-based Adaptive Self-training Semi-supervised Ensemble Learning), a novel lithology identification framework centered on Kolmogorov-Arnold Networks (KAN). This framework makes three key contributions that directly address the aforementioned challenges in lithology identification: (1) overcoming the small sample problem, (2) tackling data imbalance issues, and (3) enhancing interpretability. First, we develop a self-training semi-supervised learning strategy that effectively leverages limited labeled data alongside abundant unlabeled data to overcome the small sample problem. Our progressive thresholding mechanism gradually incorporates high-confidence predictions into the training process, substantially expanding the effective training set while maintaining classification integrity. Second, to tackle data imbalance issues, we implement a two-stage approach: applying ADASYN (Adaptive Synthetic Sampling) during the initial data preprocessing phase and again during model training to ensure balanced representation of minority lithologies. Our ensemble learning architecture combines multiple diversely initialized KAN classifiers with a meta-classifier, thereby yielding robust performance even in highly skewed lithology distributions. Third, we enhance model interpretability through a dual mechanism. At the base classifier level, KAN's learnable B-spline activation functions provide a degree of structural transparency that is absent in conventional neural networks with fixed activation functions, offering insight into how individual input features are transformed. At the system level, we integrate SHAP (SHapley Additive exPlanations) analysis as a post-hoc interpretation tool to quantitatively assess each well-logging parameter's contribution to lithology predictions across the entire ensemble. We acknowledge that the ensemble architecture (stacking and boosting) inevitably increases model complexity beyond that of a single KAN; however, the combination of KAN's interpretable components and SHAP analysis enables us to extract practical lithology identification rules, effectively bridging the gap between predictive performance and geological interpretability. The overall workflow of the proposed KASSEL framework is illustrated

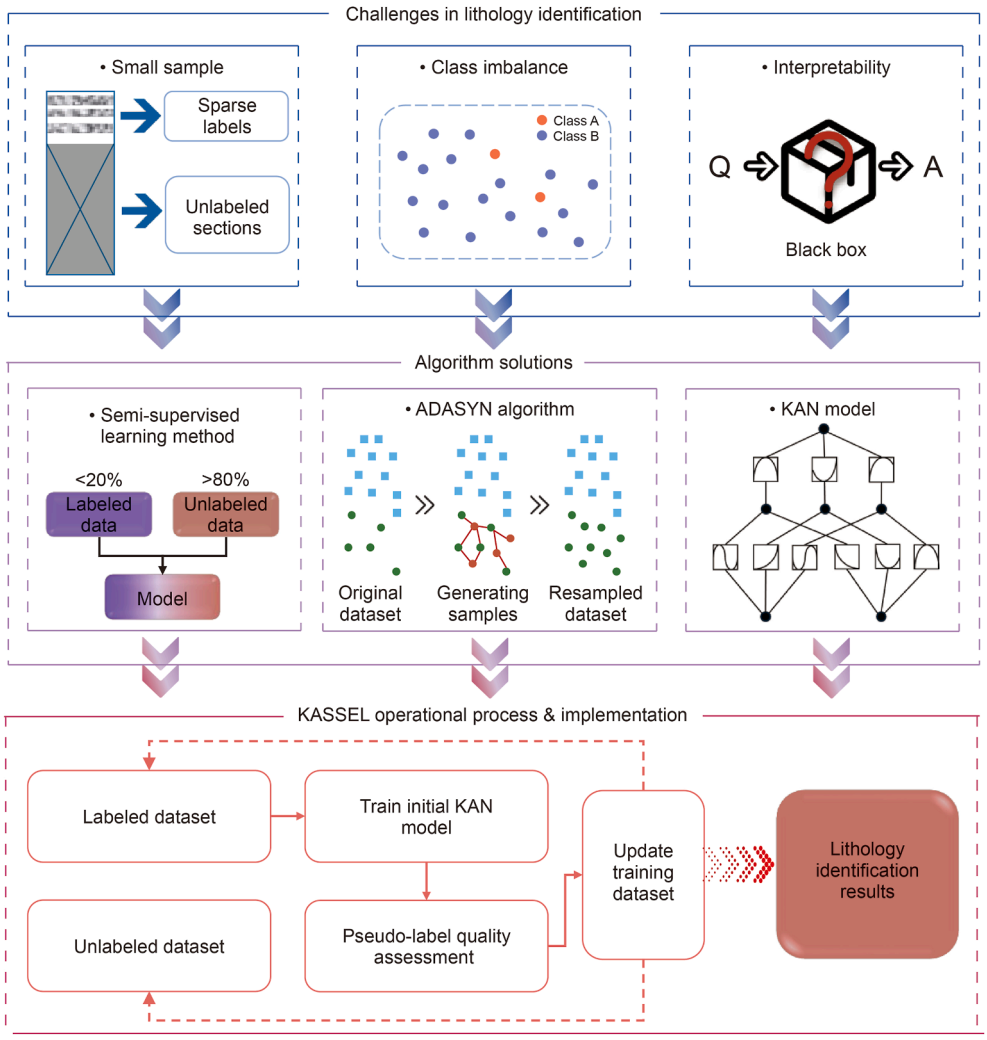


Fig. 1. Schematic representation of the proposed KASSEL (KAN-based Adaptive Self-training Semi-supervised Ensemble Learning) framework for lithology identification.

in Fig. 1, demonstrating how these components integrate to address the challenges in lithology identification. To our knowledge, this represents the first application of KAN-based self-training semi-supervised learning to lithology identification, offering a data-efficient and interpretable approach for reservoir characterization in complex geological settings.

2. Methods

The proposed KASSEL framework integrates several advanced computational techniques to address the previously discussed challenges in lithology identification. This section first introduces the key components of the framework, including Kolmogorov-Arnold Networks (KAN) as the base classifier (Section 2.1), a semi-supervised learning approach for leveraging unlabeled data (Section 2.2), and Adaptive Synthetic Sampling (ADASYN) for handling class imbalance (Section 2.3). Following these component descriptions, we present the complete architecture of the KASSEL framework (Section 2.4).

2.1. Kolmogorov-Arnold Networks (KAN)

Kolmogorov-Arnold Networks (KAN) present a novel neural network architecture centered around the Kolmogorov-Arnold theory, which states that any multivariate continuous function can be represented as a composition of continuous functions of a single variable and addition operations (Liu et al., 2024a, 2024b). This theorem, devised by Andrey Kolmogorov and refined by Vladimir Arnold, supplies the theoretical basis for KAN and can be put in mathematical terms as:

$$f(x_1, x_2, \cdots, x_d) = \sum_{q=1}^{2d+1} \Phi_q \left(\sum_{p=1}^{d} \phi_{q,p}(x_p) \right) \tag{1}$$

where $f : [0,1]^d \to \mathbb{R}$ is a continuous multivariate function, and $\phi_{q,p}$ and Φ_q are continuous univariate functions. This representation demonstrates that any continuous function of multiple variables can be constructed using only continuous functions of a single variable and the addition operation.

KAN implement this theorem through a structured neural network architecture that replaces traditional neural network activation functions with trainable B-splines. The network consists of multiple layers, each performing a specific operation aligned with the Kolmogorov-Arnold representation.

The network starts off with an input layer which receives the feature vector $\boldsymbol{X} = [x_1, x_2, \ldots, x_d]$. These inputs are then processed via a univariate function layer, where trainable univariate functions $\phi_{q,p}$ to produce intermediate outputs $z_{q,p} = \phi_{q,p}(x_p)$. The

architecture goes on with a summation layer that computes the weighted sums of these outputs, phrased as:

$$u_q = \sum_{p=1}^{d} \omega_{q,p} \cdot z_{q,p} + b_q \tag{2}$$

where $\omega_{q,p}$ are trainable weights and b_q are bias terms.

Following the summation, an outer univariate function layer applies another set of trainable univariate functions Φ_q to each sum, producing $v_q = \Phi_q(u_q)$.

Finally, the output layer generates the prediction by summing these results: $y = \sum_{q=1}^{Q} v_q$.

A distinguishing feature of KAN is the implementation of univariate functions using B-splines. These functions can be mathematically expressed as:

$$\phi(x) = \sum_{i=1}^{n} c_i B_i(x) \tag{3}$$

where $B_i(x)$ are basis spline functions and c_i are trainable coefficients. B-splines provide KAN with flexible function approximation capabilities while preserving the ability to visualize and interpret these functions after training.

The original KAN framework was mainly designed for regression tasks and doesn't inherently support classification problems. However, lithology identification is fundamentally a multi-class classification problem. To address this limitation, we extend the KAN architecture by incorporating a softmax activation function in the output layer to enable multi-class lithology identification.

The modified architecture maintains the same structure as described earlier but adds a final transformation layer. After computing the output values for each class, these values are passed via a softmax function:

$$p(y=j|x) = \frac{e^{y_i}}{\sum_{k=1}^{K} e^{y_k}} \tag{4}$$

where $p(y=j|x)$ represents the probability that the input x belongs to lithology class j, and K is the total number of lithology classes.

2.2. Semi-supervised learning

Aiming to attain higher classification accuracy levels, semi-supervised learning has proven effective in handling partially labeled data (Wang et al., 2016). The fundamental principle of semi-supervised learning involves leveraging both labeled and unlabeled data to improve classification performance beyond what would be possible using only the labeled subset. This approach involves the division of the original lithology dataset L into two distinct subsets: 1) the labeled data $\{L_{LD}\} = \{(x_\alpha, y_\alpha)|\alpha = 1, \ldots, n\}$ where x_α represents the pattern, y_α denotes the default label for x_α, with n signifies the count of labeled instances; and 2) the unlabeled data $\{L_{UD}\} = \{(x_\beta)|\beta = n+1, \ldots, n+l\}$, wherein x_β represents the pattern, and l represents the number of unlabeled instances (Vale et al., 2022).

Semi-supervised learning has demonstrated as an effective method in domains where labeled patterns are restricted (Chapelle et al., 2006). In the context of lithology identification, the semi-supervised learning algorithm can accurately predict unlabeled lithology data by learning from a limited original dataset containing lithology labels (Zhu et al., 2024).

The self-training algorithm is a straightforward semi-supervised learning technique that operates on an iterative principle (van Engelen and Hoos, 2020). The starting phase requires the construction of a classifier with a constrained subset of labeled data, which is then applied to classify the unlabeled data. Following this classification step, the classifier undergoes retraining. This iterative procedure continues until all instances within the unlabeled dataset have been incorporated.

Self-training algorithms have emerged as a promising approach for enhancing classification accuracy, particularly in scenarios where labeled data are scarce or expensive. One of their primary benefits is the ability to use small amounts of labeled data to improve the classification accuracy for large amounts of unlabeled data. The self-training algorithm creates a more extensive labeled dataset without incurring the costs associated with manual labeling through its two-step process: first iterative training on the labeled data, and second, using the resulting classifier to label the unlabeled data. Additionally, self-training algorithms can demonstrate greater robustness to noisy or outlier data points than conventional clustering algorithms. The iterative updating of the model with newly labeled data allows the algorithm to correct misclassifications and improve the classifier's overall accuracy (Zhu and Goldberg, 2009).

2.3. Adaptive synthetic sampling

The ADASYN technique, proposed by He et al. (2008), addresses the issue of imbalanced datasets by introducing an oversampling method for minority classes. Its primary objective is to mitigate the challenges posed by imbalanced class distributions. This approach involves several steps, as follows:

To determine the number of samples to synthesize for each minority sample, the ratio of majority samples in its K-nearest neighbor set is calculated. Let G represent the total synthesized sample number. For each minority sample x_i, the K-nearest neighbor set is identified in the n-dimensional space using the Euclidean distance formula, as shown in Eq. (5).

$$d = \sqrt{\sum_{i=1}^{n} (x_{Ai} - x_{Bi})^2} \tag{5}$$

Let x_{Ai} represent the i^{th} eigenvalue associated with the A^{th} data point, while x_{Bi} denotes the i^{th} eigenvalue corresponding to the B^{th} data point.

Subsequently, determine the proportion of samples assigned to the majority class among the K-nearest neighbor sets, as demonstrated by Eq. (6).

$$r_i = \Delta_i / K \tag{6}$$

The parameter K represents the number of neighboring samples while Δ_i signifies the total number of data classified as the majority class.

Eq. (7) is employed to calculate the number of synthetic samples to be generated for each minority sample x_i.

$$g_i = \frac{r_i}{\sum_i r_i} \times G \tag{7}$$

The variable r_i represents the ratio of the majority class data points in the i^{th} sample. G denotes the total number of data points required for synthesizing the minority class.

The synthesis of minority samples involves randomly selecting a sample from the minority class. This process results in the generation of synthetic data samples, as shown in Eq. (8).

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \tag{8}$$

In this context, x_i represents a minority sample, $(x_{zi}-x_i)$ denotes the difference vector in n-dimensional space.

In practical applications, KASSEL implements ADASYN with a strict protocol to prevent data leakage. Crucially, the data partitioning into training and testing sets is performed strictly prior to any oversampling. Under this protocol, ADASYN is implemented in two distinct phases utilizing only the training data. First, ADASYN is applied as a global preprocessing step on the isolated training set. This rebalances the training data distribution once before it enters the learning algorithm, ensuring a foundational balance without contaminating the test data. Second, ADASYN is incorporated locally within the algorithm structure, specifically during the ensemble learning phase. Here, localized resampling is performed dynamically on specific subsets (e.g., within cross-validation folds) of the training data. This dual-stage application allows the model to adapt to specific imbalance characteristics of different data neighborhoods while guaranteeing that the test set remains composed entirely of original, real-world samples for unbiased evaluation.

2.4. Architecture of KASSEL framework

The proposed framework, KASSEL (KAN-based Adaptive Self-training Semi-supervised Ensemble Learning), as outlined in Algorithm 1 and visualized in Fig. 2, employs modified KAN with residual connections, where batch normalization layers mitigate gradient distortion during piecewise polynomial approximation. To address class imbalance, the framework implements a two-stage adaptive synthetic sampling strategy within the training pipeline. In the first stage, after partitioning the hold-out test set, ADASYN is applied during the training data preprocessing phase. It selectively generates synthetic samples for underrepresented lithology classes while preserving the original majority class. This strictly segregates the testing data from the augmentation process. The second stage occurs within the model training process, where localized ADASYN resampling is dynamically applied to each cross-validation fold, adaptively adjusting the sampling density based on the local neighborhood distribution of minority-class instances.

The ensemble learning component implements a dual-path architecture combining stacking and boosting mechanisms. The stacking subsystem constructs a hierarchy where K independently initialized KAN models generate intermediate predictions that serve as meta-features. These meta-features are produced through stratified k-fold cross-validation, where each fold undergoes secondary ADASYN resampling to ensure balanced class representation. These meta-features are then input into an XGBoost meta-classifier. In parallel, the boosting path iteratively reweighs misclassified samples using exponential loss with updated weights.

Training optimization follows a compound strategy that integrates supervised learning with progressive pseudo-labeling. The semi-supervised component gradually incorporates unlabeled data through confidence-based thresholding, where the threshold τ is progressively reduced from τ_{start} to τ_{end}. Each base KAN undergoes adaptive learning rate scheduling with additional dynamic adjustments based on validation performance. The final ensemble model F_{final} combines the stacking meta-classifier output with weighted contributions from the boosting path models, providing robustness against overfitting even in data-limited scenarios.

Input: Labeled data X_l, Y_l; Unlabeled data X_u; Independent Test set X_{test}.
Initialize:
  //Apply ADASYN strictly on training data to prevent leakage
  Apply first-stage ADASYN: $(X_l, Y_l) \leftarrow \text{ADASYN}(X_l, Y_l)$
  Construct base KAN classifiers $\{f_k\}_{k=1}^{K}$ with:
    Residual blocks $f(x) = \text{BN}(\text{KAN}(x)) + W_{\text{shortcut}}$
    B-spline basis $\phi_g(x)$ with adaptive grid $G_{\text{new}} = \xi G_{\text{uniform}} + (1-\xi) G_{\text{data}}$
Hierarchical Ensemble Learning:
  //Stacking Phase
  Generate meta-features via k-fold cross-validation:
    $\forall k \in [1, K], \Phi_{\text{meta}} \leftarrow \underset{\text{fold}}{\cup} f_k^{(-\text{fold})}(X_{\text{val}})$
  Train meta-classifier $h_{\text{stack}} \leftarrow \text{XGBoost}(\Phi_{\text{meta}}, Y)$
  //Boosting Phase
  Initialize sample weights $\omega_i = 1/|X_l|$
  for $t = 1$ **to** T **do**:
    Train f_t with $L_{\text{total}} = \text{ECE} + \lambda_1 \|W\|_1 + \lambda_2 H(W)$
    Update weights: $\omega_i \leftarrow \omega_i \cdot \exp(\eta^{\text{I}}[f_t(x_i) \neq y_i])$
  end
Semi-supervised Learning:
  while $\tau \geq \tau_{\text{end}}$ **do:**
    Generate pseudo-labels: $\hat{Y}_u = \{x | \max f(x) > \tau\}$
    Balance dataset: $X_{\text{aug}} \leftarrow \text{ADASYN}(X_l \cup \hat{X}_u, \alpha)$
    Update classifier $F \leftarrow F + \beta f(X_{\text{aug}})$
    Adjust threshold: $\tau \leftarrow \tau - \triangle\tau_{\text{dynamic}}$
  end
Model Refinement:
  Update spline grids $G_{\text{adj}} = \text{QuantileAlign}(X, G_{\text{init}})$
  Apply adaptive learning rate: $\eta_t = \eta_0/\sqrt{1+\varepsilon t}$
return Ensemble model $F_{\text{final}} = h_{\text{stack}}(\{f_k\}_K) \oplus \sum_{t=1}^{T} \gamma_t f_t$
Output: Trained lithology classifier F

Algorithm 1. KASSEL framework for lithology identification.

3. Implementation

3.1. Study area and data preparation

The study area is located in the X block of the S tight gas reservoir, situated in the central-northern region of the Ordos Basin, China. The Ordos Basin is one of China's most important energy-producing basins, characterized as a large-scale multicycle cratonic basin with stable tectonic activity and gentle structural morphology. The target formation is the lower sub-member of the 8^{th} member of the Shihezi Formation (P_2sh^8), which serves as the primary gas-bearing interval in the S gas field. The stratigraphic units involved in this study mainly include the Shanxi and Shihezi formations, with the P_2sh^8 member being the primary focus.

The P_2sh^8 reservoir is classified as a typical tight sandstone reservoir, the reservoir lithology is predominantly composed of lithic quartz sandstone and lithic sandstone, with an average sandstone thickness of 10–18 m and an average net pay thickness of 4.8 m. Reservoir quality is generally poor, with an average porosity of 8.02% and permeability ranging from 0.01 mD to 7.547 mD (average 0.253 mD). The primary storage spaces include lithic dissolution pores, feldspar dissolution pores, intergranular dissolution pores, matrix dissolution pores, and intercrystalline pores, with minor residual intergranular pores and occasional microfractures. The reservoir pressure is 26.7 MPa with a formation

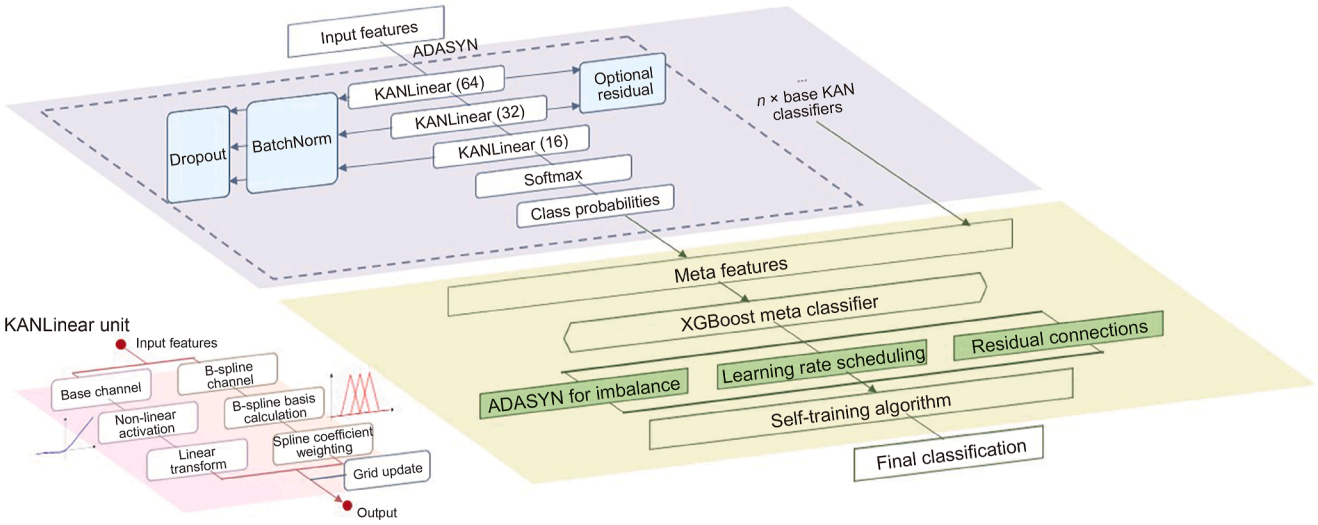


Fig. 2. Workflow of the KASSEL framework for lithology identification.

temperature of 88.5 °C, and the gas composition is dominated by methane with an average content of 92.2%.

The diverse lithological assemblage within the P_2sh^8 reservoir, including coarse sandstone (CS), fine sandstone (FS), mudstone (M), medium sandstone (MS), and pebbly sandstone (PS), along with trace amounts of argillaceous dolomite, reflects the complex multi-stage depositional history. The strong reservoir heterogeneity caused by frequent lateral facies changes, combined with the low porosity and ultra-low permeability characteristics, makes accurate lithology identification particularly challenging yet critically important for sweet spot prediction and optimal well placement in this tight gas reservoir.

In this study, 32 wells from X block of the S tight reservoir were chosen. Nine well-log parameters were employed for lithology identification, as listed in Table 1.

In this study, the precise identification of lithology types was achieved through well core experiments. A comprehensive dataset consisting of 1296 well-log readings was collected, which corresponded to the specific lithology types at respective core depths. The dataset consisted of five lithology types: coarse sandstone (CS), fine sandstone (FS), mudstone (M), medium sandstone (MS), and pebbly sandstone (PS), as summarized in Table 2. The datasets were found to be highly imbalanced, with MS and PS lithologies accounting for about 70% of the samples, while M lithology only accounts for 1.9%. This imbalance is likely to impact the effectiveness of the classifier significantly. Subsequent sections will address the adverse effects of unbalanced datasets on lithology identification tasks and the efficacy of the ADASYN algorithm in improving the performance of unbalanced lithology datasets.

This paper uses logging data and the corresponding lithology data to create the raw dataset for classification training. Moreover, conventional logging data are processed without dimensionality, and core lithology calibration is performed to optimize the logging parameters for different lithology types. As shown in Fig. 3, a radar chart analysis was employed to evaluate the discriminative capability of each logging parameter across different lithology classes, thereby guiding the dimensionality reduction of the input data. In this diagram, each axis represents a normalized logging parameter, and each lithology class forms a distinct polygon based on its mean (or median) parameter values. The degree of separation between polygons along a given axis reflects the discriminative power of the corresponding parameter. Parameters along which different lithology classes exhibit clearly separated values are considered effective for lithology discrimination, whereas parameters with heavily overlapping polygons indicate limited classification capability. The analysis results indicate that six logging parameters—SP, GR, CNL, AC, RT, and RXO—demonstrate clearly distinguishable patterns across different lithology classes, whereas the remaining three parameters—CAL, DEN, and RI—show substantial overlap among lithology polygons. Therefore, these six discriminative logging parameters are selected to establish the training dataset for lithology identification.

In this study, the wells involved are mainly exploration wells, which exhibit characteristics such as scattered locations, varied depositional environments, and heterogeneous geological structures. As a result, the samples belonging to the same lithology class demonstrate diverse distributions within the input space, adding complexity to the task of lithology identification. Fig. 4 depicts a

Table 1
Wireline logs used in this article.

Well log	Definitions	Unit
CAL	Caliper log	cm
SP	Spontaneous potential log	mV
GR	Gamma ray log	API
CNL	Compensated neutron log	%
DEN	Densilog	g/cm^3
AC	Acoustic log	μs/m
RI	Invaded zone resistivity log	Ω·m
RT	True formation resistivity log	Ω·m
RXO	Flushed zone formation resistivity log	Ω·m

Table 2
Lithologic distribution of the original dataset.

Lithology	Size	Proportion, %
CS	146	11.3
FS	195	15.0
M	25	1.9
MS	422	32.6
PS	508	39.2
All	1296	100.0

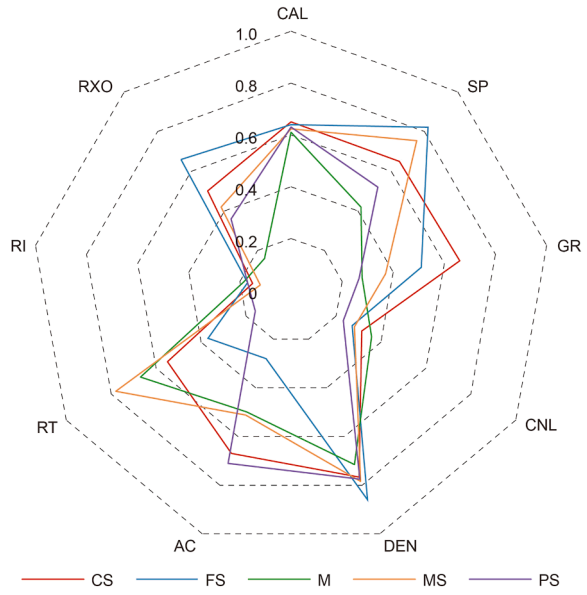


Fig. 3. Radar chart analysis of nine well logging parameters showing their discriminative patterns across different lithology classes.

cross plot of a subset of data from the dataset. The diagonal plots illustrate the kernel density estimation for each lithology in well logs, while the remaining subplots depict pairwise scatter plots of well-logging data. Notably, the data distribution of different lithologies shows considerable overlap when considering any two arbitrary well logs. This indicates that the distribution of each lithology has a substantial similarity when viewed only from the perspective of a single logging parameter. Classifying lithology using simply a single set of logging data will be difficult.

3.2. Hyperparameter optimization

To optimize the performance of ML algorithms, fine-tuning hyperparameters according to dataset characteristics is essential. Strictly prior to model development and any data augmentation, the original dataset was partitioned into training, validation, and testing sets at a ratio of 80%, 10%, and 10%, respectively. The ADASYN oversampling was subsequently applied only to the training portion, ensuring that the validation and testing sets utilized for evaluation consisted exclusively of original, non-synthetic data.

Considering the imbalanced distribution of lithology classes, a stratified sampling approach was implemented during this partitioning to maintain consistent class distribution across all subsets. For the internal evaluation, stratified five-fold cross-validation (SCV) was employed. As illustrated in the left panel of Fig. 5, the training set is randomly partitioned into five equal subsamples. In each iteration, one subsample serves as the validation set, while the remaining four are used for training (incorporating dynamic oversampling). The model's performance is evaluated based on the average accuracy across these folds. Simultaneously, the right panel of Fig. 5 visualizes the various hyperparameter combinations within the ensemble model that were explored during this process. The hyperparameter optimization was conducted using a grid search strategy (Radziwill, 2019), which systematically evaluates these parameter pairs to identify the optimal range by maximizing the average accuracy derived from the repeated SCV procedure.

The critical hyperparameters for optimization in the KASSEL framework are detailed in Table 3. In this study, Bayesian optimization was implemented through the Optuna framework to efficiently explore the hyperparameter space. The optimization process identified optimal parameter configurations that maximized the weighted F1-score across multiple evaluation rounds. Subsequently, the trained models were evaluated using the test set to select the final model. Precision (Pr), recall (Re), and F1-score were utilized in this study to evaluate the model's performance on the test set.

4. Results and discussion

4.1. Addressing class imbalance with two-stage ADASYN

A significant hurdle in lithology identification from well logs arises from class imbalance, where certain rock types are extremely underrepresented in accessible data. This section evaluates how the KASSEL framework resolves this challenge via its integrated two-stage ADASYN component. As presented in the Methods section, this strategy combines preprocessing with algorithm-level resampling to attain balanced learning for all lithology types. The ensuing experiments demonstrate the validity of this strategy in improving the performance of classification, especially for minority lithology classes suffering from poor recognition rates.

4.1.1. Impact of class imbalance on classification performance

As described in the Methods section, this strategy combines preprocessing with algorithm-level resampling to achieve balanced learning for all lithology types. Table 4 presents a comparison of classification performance metrics before and after applying the two-stage ADASYN algorithm.

Table 4 demonstrates significant improvements in classification performance following the implementation of the two-stage ADASYN algorithm within our KASSEL framework. Most notably, the minority mudstone (M) class, which completely failed in classification using the original imbalanced dataset (precision, recall, and F1-score all 0.0), achieved exceptional performance after balancing (precision 0.83, recall 1.00, F1-score 0.91). Similarly, other classes show substantial improvements, with the coarse sandstone (CS) class achieving perfect precision (1.00) and near-perfect recall (0.93). Overall metrics demonstrate the comprehensive impact of our approach, with macro-average precision improving from 0.64 to 0.88. These results clearly demonstrate the KASSEL framework's effectiveness in addressing class imbalance challenges.

To further investigate the classification performance improvements, we analyzed the confusion matrices before and after implementing our two-stage ADASYN algorithm, as shown in Fig. 6.

In the original imbalanced dataset (Fig. 6(a)), the minority class M demonstrates no correct classifications. Most M samples (60%) are incorrectly classified as PS, while the remaining samples are misclassified as CS (20%) and MS (20%). This demonstrates the severe impact of class imbalance on minority class detection. The CS and FS classes also show moderate performance, with correct classification rates of only 0.62 and 0.67, respectively. CS samples are frequently misclassified as MS (17%) and PS (20%), while FS samples show similar confusion with mud-containing lithologies.

Following the implementation of our two-stage ADASYN algorithm (Fig. 6(b)), we observed significant improvements in classification performance. The minority class M achieves perfect classification (1.00 accuracy), effectively resolving the previous inability to detect this lithology type. Similarly, the FS class reaches 1.00 correct classification, eliminating all previous

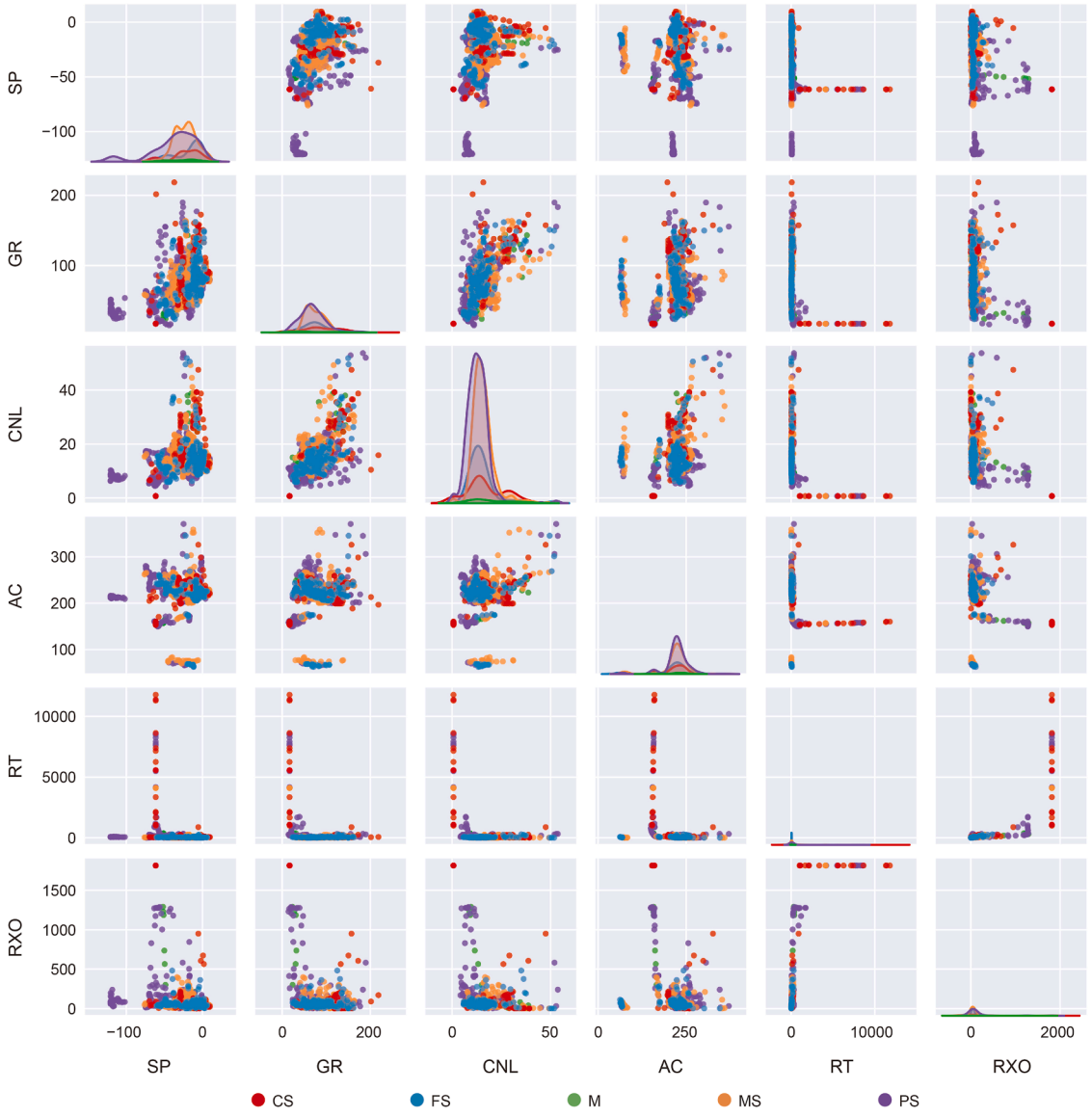


Fig. 4. Crossplot of the dataset.

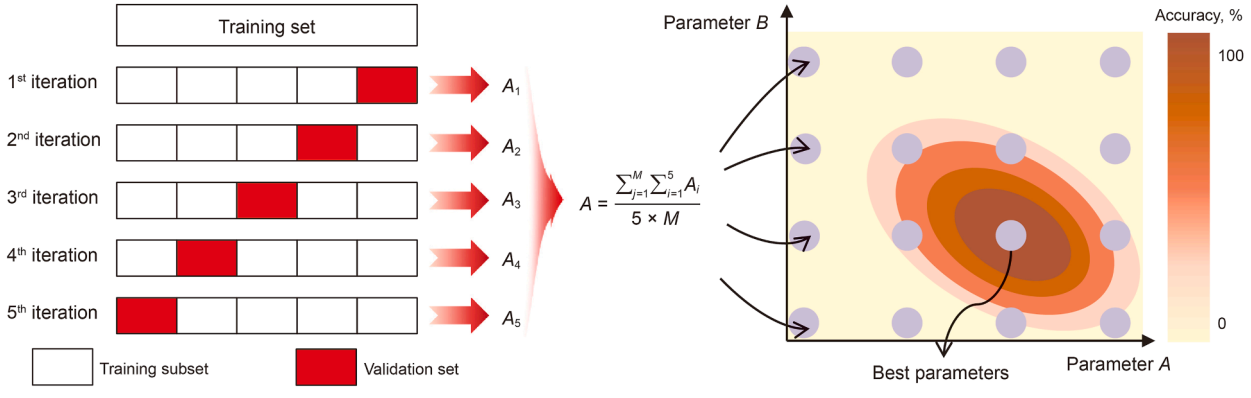


Fig. 5. Schematic diagram of SCV and grid search method.

Table 3
Hyperparameters optimization results.

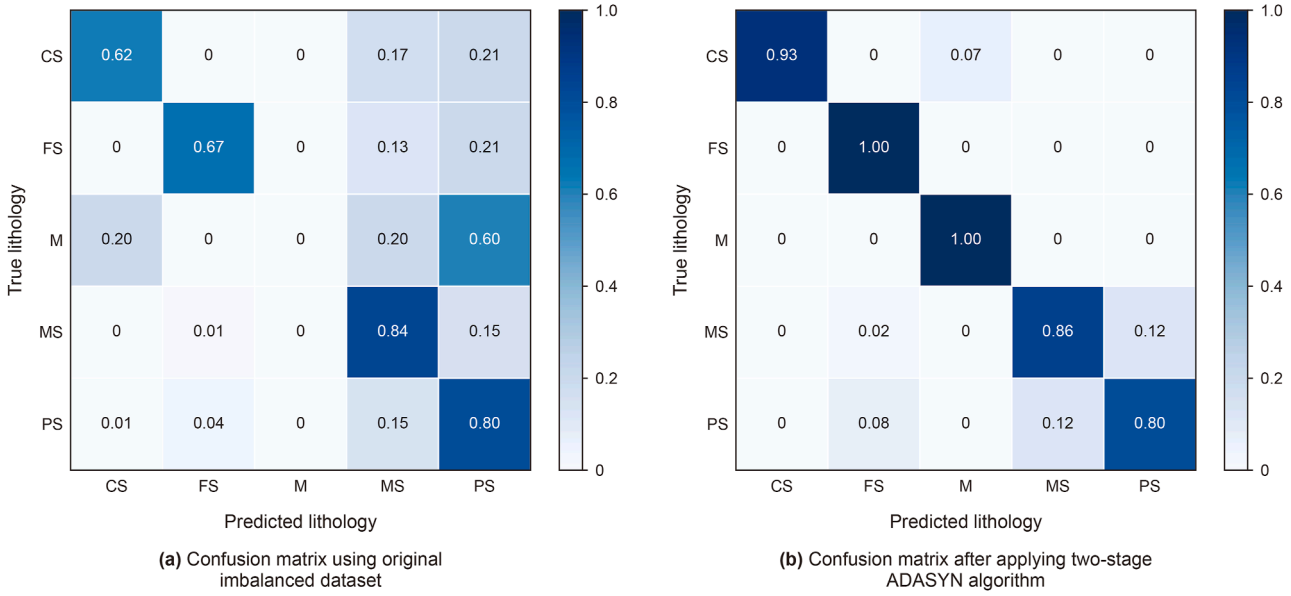
Algorithm component	Hyperparameters	Definitions	Parameter search intervals	Best parameter configurations
KAN base model	*n*_base_estimators	Number of base KAN classifiers in the ensemble	3–10	3
	hidden_layers	Structure and number of neurons in hidden layers	Layers: 1–4 Units per layer: 32–512	Layers: 1 Units per layer:406
	activation	Type of activation function used in the neural network	SiLU, ReLU, LeakyReLU, tanh	tanh
	dropout_rate	Proportion of randomly dropped neurons during training to prevent overfitting	0.1–0.6	0.247
XGBoost meta-classifier	*n*_estimators	Number of trees generated in the boosting iterations	50–300	64
	learning_rate	Contribution of each tree to the final result (shrinkage rate)	0.01–0.3	0.003
	max_depth	Maximum depth of trees, controlling model complexity	2–8	5
	subsample	Sampling ratio of training instances	0.6–1.0	0.535
	colsample_bytree	Feature sampling ratio when building trees	0.6–1.0	0.509
	min_child_weight	Minimum sum of instance weight needed in a child node	1–10	8
	gamma	Minimum loss reduction required for node splitting (regularization parameter)	0–5	0.650
Self-training algorithm	τ_{start}	Initial confidence threshold for pseudo-label selection	0.7–0.99	0.9
	τ_{end}	Final confidence threshold in the last self-training iteration	0.5–0.9	0.7
	τ_{decay}	Rate at which threshold decreases from start to end value	Linear, exponential, step-wise	Linear

Table 4
Comparison of classification performance metrics before and after applying two-stage ADASYN algorithm.

	Original imbalanced dataset			After two-stage ADASYN		
Lithology	Precision	Recall	F1-score	Precision	Recall	F1-score
CS	0.90	0.62	0.73	1.00	0.93	0.97
FS	0.81	0.67	0.73	0.77	1.00	0.87
M	0.00	0.00	0.00	0.83	1.00	0.91
MS	0.76	0.84	0.79	0.86	0.88	0.87
PS	0.74	0.82	0.78	0.91	0.78	0.84
Macro avg	0.64	0.59	0.61	0.88	0.92	0.89
Weighted avg	0.76	0.77	0.76	0.88	0.87	0.87

misclassifications. The CS class shows substantial improvement from 0.62 to 0.93 in accuracy.

The comparison of confusion matrices clearly demonstrates the effectiveness of our two-stage ADASYN approach in addressing classification challenges, particularly for minority classes. A noteworthy observation is the perfect recall (1.00) achieved by the Mudstone (M) class after balancing. While this result may raise concerns about potential overfitting, it should be interpreted in the context of the extremely small sample size of this class. The M class constitutes only 1.9% of the original dataset (25 samples), resulting in approximately 2–3 test samples after stratified partitioning. With such a limited test subset, classification outcomes for this class are inherently discrete—even a single misclassification would cause a substantial drop in accuracy. The reported results represent the average performance obtained through stratified five-fold cross-validation, which mitigates the risk of evaluation bias from any single data split. Nevertheless, we acknowledge that the evaluation of rare lithology classes would benefit from larger datasets in future studies, and the incorporation of pseudo-label quality monitoring mechanisms represents a promising direction for further improving the robustness of the self-training process.



(a) Confusion matrix using original imbalanced dataset

(b) Confusion matrix after applying two-stage ADASYN algorithm

Fig. 6. Confusion matrices comparing lithology identification performance before and after applying the two-stage ADASYN algorithm. **(a)** Confusion matrix using original imbalanced dataset; **(b)** confusion matrix after applying two-stage ADASYN algorithm.

4.1.2. Balancing lithology data using ADASYN algorithm

To examine how the KASSEL framework's data preprocessing stage transforms the original imbalanced dataset, we analyzed the distribution changes resulting from the ADASYN algorithm implementation. The ADASYN algorithm utilizes a 'not majority' sampling strategy, which resamples all classes except the majority class. This approach led to oversampling of the four lithology classes in the original dataset, excluding the majority class of PS lithology. The size and distribution of the dataset after balance are shown in Table 5.

The balanced lithology dataset demonstrates a more uniform distribution following the resampling of minority classes, generating additional synthetic samples. The increase in sample size varies according to the original data volume for each lithology type. For example, the MS lithology, which contained the second largest number of samples in the original dataset, increased by only two samples, while the M lithology, which contained the smallest number, increased by 236 samples. The number of lithologic data for CS, FS, M, and MS increased to varying degrees, making the number of the minority classes relatively balanced with that of the majority classes compared to the original dataset (Table 2). To assess the quality of the samples generated by the ADASYN algorithm, we conducted a visual analysis using the t-SNE algorithm, which mapped the multidimensional input well logs to three-dimensional coordinates (x_1, x_2, and x_3) for each data subset observation. The data distribution of the original lithology dataset is presented in Fig. 7(a). The original dataset demonstrates characteristics of the curse of dimensionality, with data points primarily distributed at the space boundaries rather than within the interior. In addition, the number of M lithologic data points is significantly lower than that of other lithologic data points.

The data distribution of the dataset following the ADASYN algorithm implementation is presented in Fig. 7(b). The ADASYN algorithm utilizes the density distribution as a criterion to determine the number of synthetic samples generated for each minority data instance. A comparison of Fig. 7(a) and (b) demonstrates that the ADASYN algorithm generates additional synthetic samples in regions with lower original data point density, thereby increasing the number of minority samples (darker data points).

4.1.3. Comparison of classification accuracies across different processing methods

In this section, we assess the impact of various imbalance mitigation strategies on lithology identification accuracy. Fig. 8 displays the classification accuracies across all lithology types under four distinct processing scenarios.

Our analysis investigates four processing approaches: (1) No Processing, where the original imbalanced dataset is used directly; (2) Preprocessing, where ADASYN is applied before model training as described in Section 4.1.2; (3) Algorithm Integration, where ADASYN resampling is dynamically applied within each cross-validation fold; and (4) Combined Methods, which implements both preprocessing and algorithm integration strategies.

Table 5
The lithology distribution of the dataset processed by the ADASYN algorithm.

Lithology	Size	Proportion, %
CS	342	18.0
FS	313	16.5
M	311	16.4
MS	424	22.3
PS	508	26.8
All	1898	100.0

The results demonstrate significant variations in classification performance depending on the specific imbalance mitigation strategy employed. Most notably, the mudstone (M) class shows remarkable improvement with suitable processing. Without any processing, the M class exhibits a classification accuracy of 0. Both the Preprocessing and Algorithm Integration methods effectively address this issue, raising the accuracy to 1.00. The Combined Methods approach attains an accuracy of 0.83 for class M, demonstrating a substantial boost compared to No Processing, but a little less effective than the individual techniques.

An interesting finding is the trade-off between minority- and majority-class performance. While the Preprocessing approach significantly increases the recognition of minority classes, it simultaneously reduces the classification accuracy for majority classes compared to No Processing.

The Algorithm Integration method demonstrates more balanced improvements, achieving excellent minority class recognition while maintaining reasonable performance for majority classes. This suggests that dynamic, localized resampling during model training provides a more adaptable approach for addressing class imbalance.

The Combined Methods strategy achieves the maximum overall accuracy (0.88) and yields the most balanced performance across all classes. It particularly excels in its application to coarse sandstone (CS), achieving perfect classification (1.00), and significantly improves the performance of medium sandstone (MS, 0.90) and pebbly sandstone (PS, 0.88). It is important to note that the two-stage ADASYN strategy is not a simple repetition of the same operation. The two stages serve fundamentally different purposes and operate at different scopes. The first stage (Preprocessing) provides a global rebalancing of the entire training set before model training begins, establishing a foundational balance across all lithology classes. The second stage (Algorithm Integration) performs localized, dynamic resampling within each cross-validation fold during the ensemble training process, adaptively adjusting the sampling density based on the specific neighborhood distribution encountered in each fold. As evidenced by the results in Fig. 8, this combined strategy achieves the highest overall accuracy (0.88) with the most balanced performance across all classes, whereas each individual stage alone shows trade-offs: Preprocessing significantly improves minority class recognition but reduces majority class accuracy, while Algorithm Integration achieves more balanced improvements but does not fully optimize minority class performance. The complementary nature of these two stages-global structural balancing followed by local adaptive refinement-explains why the combined approach outperforms either stage applied independently. These findings demonstrate that although all imbalance mitigation strategies increase overall classification accuracy compared to No Processing, their impacts vary significantly across different lithology classes.

4.2. Handling limited labeled data through semi-supervised learning

The scarcity of labeled data presents one of the most significant challenges in the characterization of unconventional reservoirs. This section evaluates the KASSEL framework's ability to address scenarios featuring scarce labeled samples, directly addressing the small-sample challenge mentioned in the introduction. In real-world well logging interpretation activities, cost and technical limitations typically render comprehensive core-based lithology labeling unavailable, and small labeled datasets become the norm rather than the exception. The semi-supervised learning module of KASSEL is specifically designed to take advantage of unlabeled data

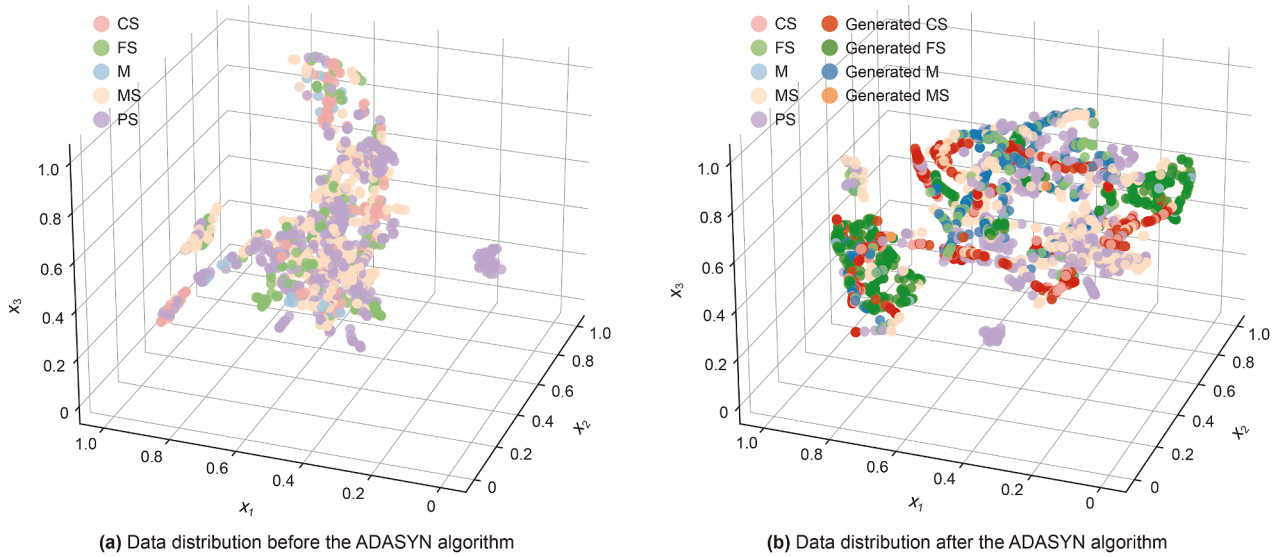


Fig. 7. t-SNE visualization plot of the database. **(a)** Data distribution before the ADASYN algorithm; **(b)** data distribution after the ADASYN algorithm.

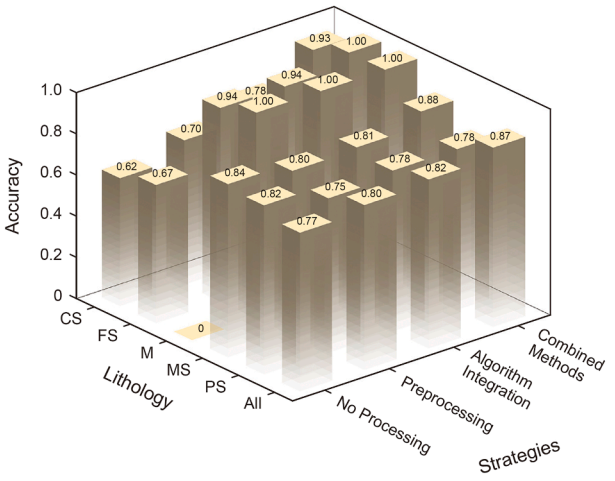


Fig. 8. Three-dimensional bar chart illustrating classification accuracies for different lithology types across various data processing methods.

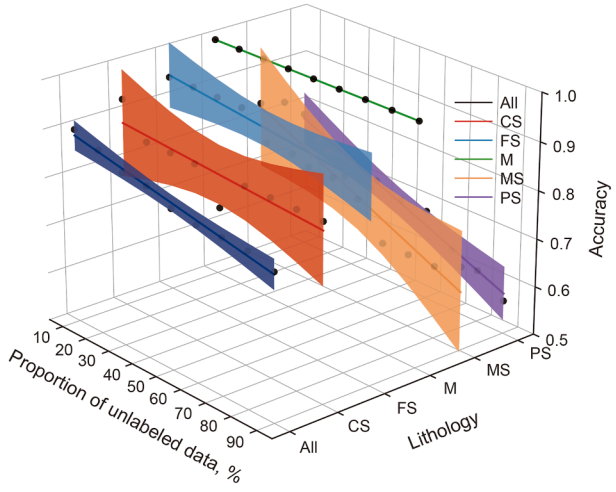


Fig. 9. Classification accuracy comparison of KASSEL under different proportions of unlabeled data.

along with limited labeled samples, allowing the framework to sustain robust performance even under harsh labeling conditions. To mimic these conditions, we systematically altered the amount of unlabeled data as a proportion from 10% to 90% of the total dataset.

4.2.1. Classification performance of KASSEL with limited labels

While addressing class imbalance is crucial, another equally significant challenge in unconventional reservoir characterization is the scarcity of labeled data. The classification performance of KASSEL under varying proportions of labeled data, ranging from 90% down to 10%, is presented in Fig. 9. A primary challenge in self-training frameworks is the risk of error propagation, where incorrect pseudo-labels are iteratively added to the training set and degrade model performance. To address this issue, our framework employs a strict progressive thresholding mechanism. This mechanism ensures that only high-confidence predictions from the KAN ensemble are incorporated into the training set, thereby preserving classification integrity even as the availability of labeled data becomes increasingly scarce.

As observed in Fig. 9, classification accuracy decreases at varying rates across different lithology classes as the proportion of unlabeled data increases from 10% to 90%. The classes of CS, FS, and M demonstrate remarkable stability; notably, the M class maintains near-perfect accuracy even when 90% of the data is unlabeled. Conversely, classes MS and PS exhibit a more substantial decline, with class PS showing the most pronounced reduction. However, even under the extreme condition where 90% of samples lack lithology labels, the KASSEL framework achieves a robust macro-average accuracy. This resilience demonstrates the framework's effectiveness in leveraging the intrinsic structure of unlabeled data through its ensemble architecture, effectively countering the label scarcity problem.

4.2.2. Comparative analysis with alternative classification methods

In this section, we conduct a comprehensive comparative evaluation to validate the overall performance of KASSEL and the specific contribution of its components. We compare KASSEL (KAN-ADASYN-Self-training) against five alternative approaches,

including two ablation variants and three established deep learning baselines. The ablation variants are MLP-ADASYN-Self-training, which replaces KAN with a standard Multilayer Perceptron (MLP) to test the necessity of the KAN architecture, and KAN, which utilizes KAN without self-training or ensemble strategies to verify the contribution of the semi-supervised ensemble mechanism. The deep learning baselines include Transformer, Long Short-Term Memory network (LSTM), and Convolutional Neural Network (CNN), representing state-of-the-art architectures that have been widely adopted in well-log analysis and geoscience applications. All baseline models were trained under identical data partitioning and evaluation protocols to ensure a fair comparison. The classification performance of these six models under unlabeled data proportions ranging from 10% to 90% is illustrated in Fig. 10.

The results in Fig. 10 reveal several important findings. KASSEL consistently achieves the highest classification accuracy across all tested unlabeled data proportions, ranging from 0.87 at 10% missing labels to 0.80 at 90%, with a total decline of only 8.2%.

Among the deep learning baselines, the Transformer model performs best, achieving a comparable accuracy of 0.87 at 10% unlabeled data. However, its performance drops sharply to 0.70 at 90%—a decline of approximately 20%. The LSTM and CNN models show even more pronounced degradation, declining to 0.55 and 0.57 at 90%, respectively. These results highlight a critical limitation of standard deep learning architectures: while they achieve competitive performance with sufficient labeled data, they lack mechanisms to leverage unlabeled samples effectively and are therefore highly sensitive to label scarcity.

The ablation analysis further elucidates the contribution of individual components. KASSEL consistently outperforms its MLP-based variant by approximately 4–5 percentage points, confirming that KAN's learnable B-spline activation functions capture complex non-linear lithological features more effectively than fixed activation functions in standard MLPs. Meanwhile, the KAN baseline exhibits the most severe degradation (from 0.74 to 0.60), underscoring that the KAN architecture alone is insufficient to handle extreme label scarcity without the self-training semi-supervised ensemble mechanism.

4.3. Enhancing model interpretability with SHAP analysis

To address the interpretability limitations of traditional lithology identification methods, we integrated SHAP (SHapley Additive exPlanations) analysis into our KASSEL framework. While the ensemble architecture of KASSEL enhances predictive performance, it inevitably introduces structural complexity that limits direct model inspection. SHAP analysis serves as the primary interpretability tool by quantifying the contribution of distinct well-logging parameters to prediction accuracy based on the fully trained model, providing valuable insights into the model's decision-making mechanism (Lundberg and Lee, 2017). This post-hoc approach complements the inherent transparency of KAN's learnable B-spline activation functions at the base classifier level. The SHAP analysis was conducted on our trained KASSEL framework using KernelExplainer with optimized background and sample sizes (Fig. 11). Notably, the SHAP-derived feature importance ranking is broadly consistent with the preliminary screening results obtained from the spider web diagram analysis in Section 3.1, which mutually validates both approaches.

Our analysis revealed that spontaneous potential (SP) is the most influential parameter, with an average feature importance of 0.071 across all lithologies and especially high importance values for sandstone varieties (0.111 for MS and 0.107 for PS). Regarding the overall importance, acoustic transit time (AC) and neutron porosity (CNL) came second and third, with average importance values of 0.052 and 0.049, respectively. Different lithological formations showed distinct sensitivity patterns. MS and PS exhibited the highest overall sensitivity to well-logging parameters (total sensitivity scores of 0.479 and 0.396, respectively), whereas M showed the lowest (total sensitivity of 0.126), with gamma ray (GR) being the dominant parameter for M identification (importance value of 0.042). Various lithology groups displayed unique sensitivity patterns, with MS showing particularly high sensitivity to SP (0.111) and AC (0.092), while CS demonstrated more balanced sensitivity across all parameters. These findings are consistent with established geological principles, as SP effectively captures formation fluid properties and permeability variations that are critical for distinguishing between rock types, while GR's sensitivity to clay content makes it particularly valuable for M identification despite the overall lower sensitivity of this lithology type to logging parameters in general.

By quantifying each parameter's contribution to lithology identification, our KASSEL framework combined with SHAP analysis transforms from a "black box" into an interpretable system that enhances both the reliability of classifications and their geological significance.

The beeswarm plots in Fig. 12 illustrate the detailed visualization of how individual well-logging measurements influence lithology identification across different samples. Each point represents a single sample, with color indicating feature value (red for high, blue for low) and horizontal position showing the SHAP value impact (positive values increase the probability of a specific lithology, whereas negative values decrease it).

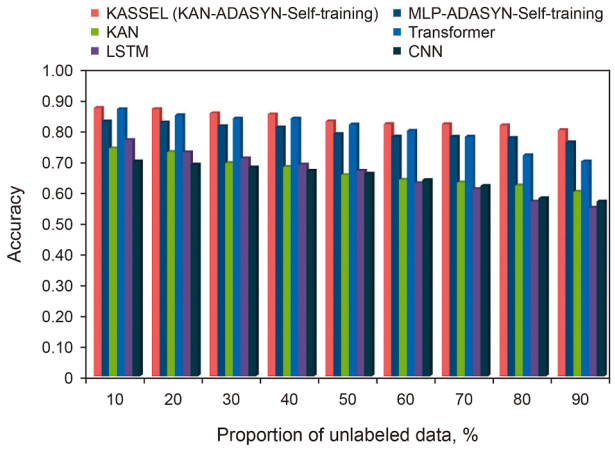


Fig. 10. Classification accuracy comparison of KASSEL with other models/algorithms.

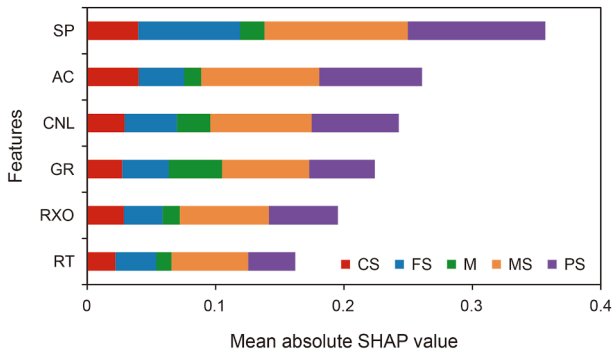


Fig. 11. Mean absolute SHAP values of input features across different lithology classes.

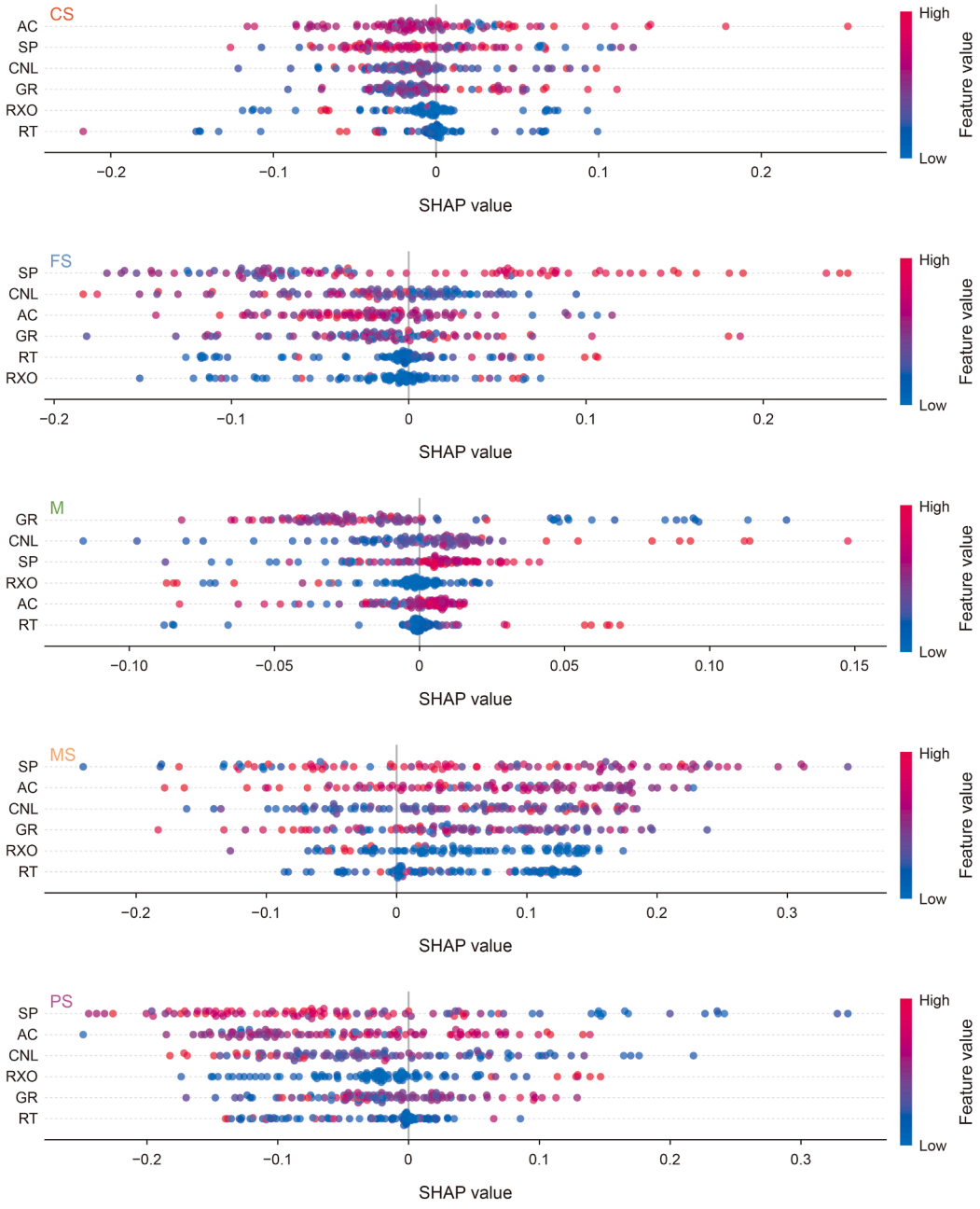


Fig. 12. SHAP beeswarm plots revealing feature value impact distribution across lithology classes.

For CS, acoustic transit time (AC) and spontaneous potential (SP) show the widest distribution of SHAP values, with high AC values generally having a positive impact on classification. This reflects the distinctive acoustic properties of coarse sandstone compared to other lithologies.

For FS, SP measurements demonstrate significant influence with a broad distribution pattern. Higher SP values typically contribute positively to FS identification, while CNL also shows considerable impact, reflecting the importance of porosity measurements for this lithology.

M displays a notably different pattern, with gamma ray (GR) exhibiting the strongest influence. High GR values consistently show positive SHAP contributions, confirming the well-established relationship between gamma radiation and clay content characteristic of mudstone. The overall narrower SHAP value range for mudstone indicates its more distinctive logging signature compared to other lithologies.

MS class reveals the most extensive SHAP value range among all lithologies, with SP and AC showing particularly strong positive contributions at high feature values. This wider distribution reflects the complex petrophysical properties of MS and its sensitivity to multiple logging parameters.

For PS, SP again demonstrates significant impact, though with a more balanced distribution between positive and negative contributions. The distribution patterns for PS show similarities to MS but with distinctive differences in parameters like AC and CNL.

These distribution patterns complement our previous findings on mean absolute SHAP values, providing deeper insights into how specific parameter values influence classification outcomes across the entire dataset.

4.4. Field application

In this section, we demonstrate the effectiveness of the proposed KASSEL framework through its application to lithology identification in the S tight reservoir. The S field is located in the central and northern parts of the Ordos Basin, with particular focus on the main gas-bearing formation, the P_2sh^8 member. This formation is crucial for lithology identification within the S gas field. Notably, early production analysis of the gas reservoir demonstrates the significant influence of lithology and physical properties on reservoir quality. The reservoir's complex and anisotropic nature renders traditional lithology interpretation methods inadequate.

The lithology identification task involved Well Y in the target block, with the target formation (P_2sh^8) located between 2638 m and 3076 m. Due to incomplete core recovery, the well lacks lithology labels for over 50% of the interval. This scenario presents an ideal test case for KASSEL's ability to predict lithology in uncored sections. The KASSEL algorithm is used to predict the lithology classes, and the real core experiment interpretation data is used to verify the prediction accuracy. As shown in Fig. 13, the algorithm demonstrates high classification accuracy for lithology identification (Litho ID), with good agreement between the existing data segments and the experimental data. The model also achieves high-precision classification results for the minority lithology classes.

A key advantage of the proposed KASSEL framework is its interpretability, facilitated through integrated SHAP analysis. This analysis quantifies the contribution of each logging parameter to lithology identification decisions, enabling the extraction of practical classification standards. The SHAP analysis revealed distinctive logging parameter ranges that influence the identification of specific lithologies. Table 6 presents these classification standards, highlighting the ranges of well-log parameters that serve as indicators for each lithology type.

For practical field application, we have synthesized these findings into a simplified interpretative framework that emphasizes the most discriminative features for each lithology. Coarse sandstone (CS) is characterized by moderate to low SP values (−61 to −11 mV), moderately high gamma ray readings (72 to 147 API), and high acoustic transit time (231 to 263 μs/m). Fine sandstone (FS) exhibits relatively high SP values (−12 to 1 mV), high gamma ray readings (92 to 152 API), distinctively low neutron porosity (9% to 12%), and low to moderate acoustic transit time (114 to 218 μs/m). Mudstone (M), often challenging to discriminate using conventional methods, shows a distinctive signature with high gamma ray values (75 to 105 API) and high neutron porosity (26% to 31%), while most other logging parameters offer minimal discriminative value. Medium sandstone (MS) presents moderate SP values (−38 to −7 mV), moderate to low gamma ray readings (48 to 91 API), and intermediate acoustic transit time (161 to 228 μs/m). Finally, pebbly sandstone (PS) is identifiable by low SP values (−74 to −39 mV), moderately high gamma ray readings (86 to 120 API), low to moderate neutron porosity (1% to 15%), high acoustic transit time (227 to 269 μs/m), and distinctively high shallow resistivity values (143 to — Ω·m).

These empirically derived classification standards constitute a significant practical contribution to this research. By translating complex machine learning outcomes into explicit logging parameter ranges, our approach bridges the gap between advanced computational methodologies and daily operational requirements in reservoir characterization. Field engineers can apply these standards directly to interpret lithology in uncored intervals or new wells, reducing dependence on expensive and

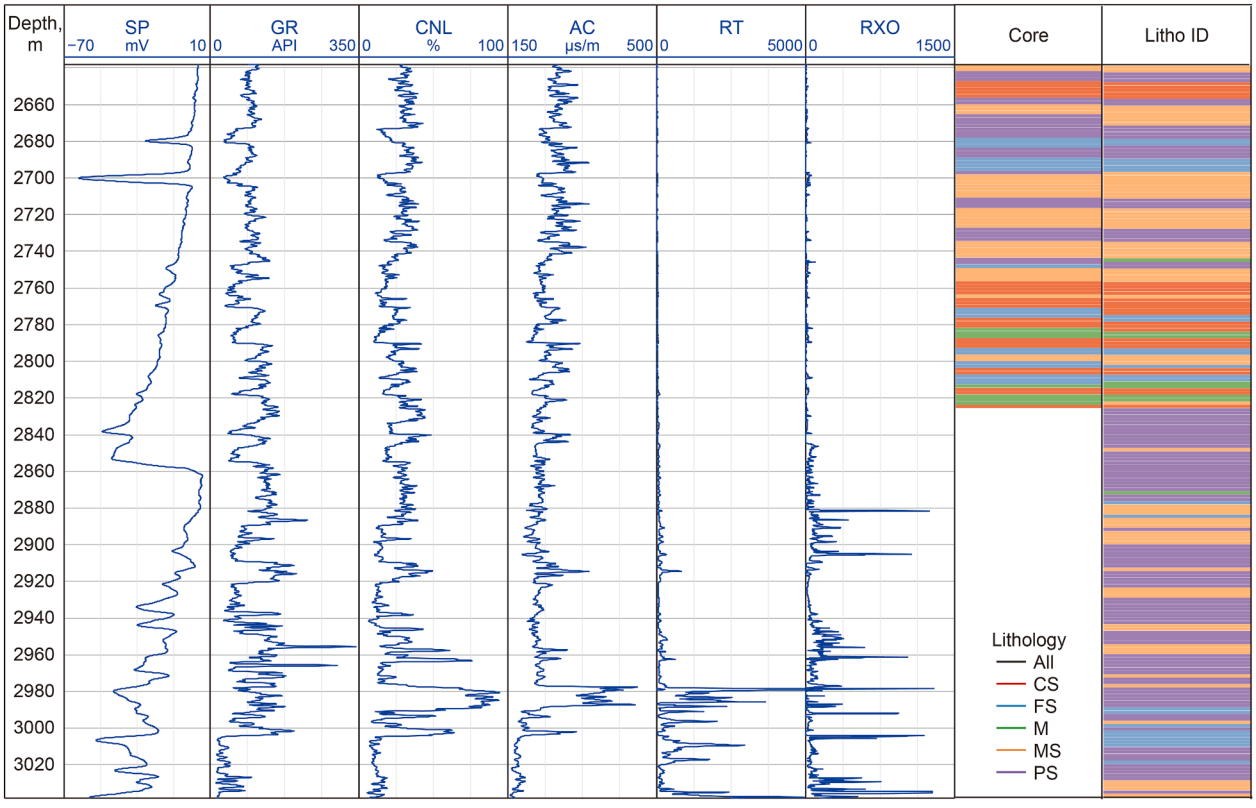


Fig. 13. Lithology identification results of the KASSEL framework.

Table 6
Lithology identification standards derived from SHAP analysis of the KASSEL framework.

Lithology	SP, mV	GR, API	CNL, %	AC, μs/m	RT, Ω·m	RXO, Ω·m
CS	−61.35 to −10.59	71.79 to 146.64	12.12 to 29.79	230.84 to 263.16	30.95 to 77.84	45.31 to 106.34
FS	−11.96 to 1.13	92.17 to 152.31	9.27 to 11.91	113.58 to 217.64	39.62 to —	42.21 to —
M	—	74.88 to 104.88	—	25.90 to 31.37	—	—
MS	−37.50 to −7.43	48.28 to 91.22	11.12 to 27.48	161.19 to 228.17	23.81 to 84.55	26.30 to 78.49
PS	−74.39 to −38.73	86.19 to 120.22	1.11 to 15.12	227.25 to 269.18	—	142.98 to —

time-consuming core analysis while maintaining classification accuracy.

It should be noted that the quantitative classification standards presented in Table 6 are empirically derived from the specific geological and petrophysical conditions of the S tight reservoir in the Ordos Basin. These numerical thresholds reflect the particular depositional environment, diagenetic history, and fluid properties of this formation, and should not be directly applied to other reservoirs without recalibration. However, the KASSEL framework itself, including its methodology for integrating semi-supervised learning, adaptive resampling, and SHAP-based interpretation, is designed to be transferable. When applied to a new reservoir or basin, the framework can be retrained with local well-log and core data to derive region-specific classification standards. In this sense, the generalizable contribution of this work lies in the workflow and methodology rather than in the specific numerical ranges, which should be treated as case-specific reference values.

5. Conclusions

This study proposes KASSEL, a novel KAN-based adaptive self-training semi-supervised ensemble learning framework, to address the critical challenges of lithology identification in unconventional reservoirs. By employing Kolmogorov-Arnold Networks as base classifiers within an ensemble architecture, the framework effectively captures complex nonlinear relationships in well-log data. Experimental results demonstrate that KASSEL achieves a macro F1-score of 0.89 on the test dataset, significantly outperforming traditional machine learning and deep learning approaches. The primary conclusions are summarized as follows.

(1) The proposed two-stage ADASYN strategy successfully mitigates the class imbalance problem inherent in geological data. This strategy improves the macro-average F1-score from 0.61 to 0.89. Most notably, the previously undetectable minority mudstone class is successfully recovered, demonstrating the framework's capability to recognize rare lithologies without compromising majority class performance.

(2) The self-training semi-supervised learning component effectively addresses label scarcity. By iteratively incorporating high-confidence predictions from unlabeled data, the framework significantly expands the effective training set, ensuring robust performance even when labeled core samples are extremely limited, as is common in early-stage exploration.

(3) The framework offers enhanced interpretability compared to conventional "black-box" models. Although the ensemble architecture increases structural complexity, combining SHAP-based post-hoc analysis with KAN's learnable B-spline activation functions provides a practical interpretability pathway. This approach identifies Spontaneous Potential (SP) and Acoustic Transit Time (AC) as the most influential parameters for lithology discrimination and enables the extraction of explicit classification standards for field application.

(4) The KASSEL framework establishes a generic and transferable solution for complex pattern recognition tasks. Its successful application in the S tight reservoir demonstrates the effectiveness of combining interpretable deep learning with robust imbalance-handling strategies. This framework holds significant potential for extension to other scientific and engineering domains characterized by high-dimensional data and label scarcity.

CRediT authorship contribution statement

Yu-Xuan Deng: Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization. **Wen-Dong Wang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Feng Jia:** Writing – review & editing, Methodology, Formal analysis. **Yu-Liang Su:** Supervision, Resources, Project administration. **Xin-Yu Zhuang:** Visualization, Validation, Software. **Cun-Qi Jia:** Writing – review & editing, Investigation. **Yong-Mao Hao:** Validation, Data curation. **Lei Li:** Investigation, Formal analysis.

Data availability

Data will be made available on request.

Funding

This study was supported by the National Science and Technology Major Project of China (No. 2026ZD1404005) and the National Natural Science Foundation of China (52274056 and 52374064).

Declaration of competing interest

The authors declare no conflict of interest regarding the publication of this paper.

References

Al-Mudhafar, W.J., 2017. Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. J. Pet. Explor. Prod. Technol. 7 (4), 1023–1033. https://doi.org/10.1007/s13202-017-0360-0.

Ali, M., Zhu, P., Ma, H., et al., 2025. Data-driven machine learning approaches for precise lithofacies identification in complex geological environments. Geo Spatial Inf. Sci. 28 (4), 1722–1742. https://doi.org/10.1080/10095020.2024.2405635.

Barach, B.A.B., Jaafar, M.Z., Gaafar, G.R., et al., 2021. Development and identification of petrophysical rock types for effective reservoir characterization: Case study of the Kristine Field, offshore Sabah. Nat. Resour. Res. 30 (3), 2497–2511. https://doi.org/10.1007/s11053-021-09851-3.

Bhattacharya, S., Carr, T.R., Pal, M., 2016. 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. 33, 1119–1133. https://doi.org/10.1016/j.jngse.2016.04.055.

Chang, H.C., Kopaska-Merkel, D.C., Chen, H.C., et al., 2000. Lithofacies identification using multiple adaptive resonance theory neural networks and group decision expert system. Comput. Geosci. 26 (5), 591–601. https://doi.org/10.1016/S0098-3004(00)00010-8.

Chapelle, O., Schölkopf, B., Zien, A., 2006. A discussion of semi-supervised learning and transduction. In: Chapelle, O., Schölkopf, B., Zien, A. (Eds.), Semi-Supervised Learning. The MIT Press, Cambridge, MA, pp. 473–478. https://doi.org/10.7551/mitpress/9780262033589.003.0025.

Dong, S., Yang, X., Xu, T., et al., 2024. Semi-supervised neural network for complex lithofacies identification using well logs. IEEE Trans. Geosci. Rem. Sens. 62, 5926519. https://doi.org/10.1109/TGRS.2024.3450103.

Dong, S., Zeng, L., Du, X., et al., 2022. Lithofacies identification in carbonate reservoirs by multiple kernel fisher discriminant analysis using conventional well logs: A case study in A oilfield, Zagros Basin, Iraq. J. Petrol. Sci. Eng. 210, 110081. https://doi.org/10.1016/j.petrol.2021.110081.

Dong, S.Q., Sun, Y.M., Xu, T., et al., 2023b. How to improve machine learning models for lithofacies identification by practical and novel ensemble strategy and principles. Pet. Sci. 20 (2), 733–752. https://doi.org/10.1016/j.petsci.2022.09.006.

Dong, S.Q., Zhong, Z.H., Cui, X.H., et al., 2023a. A deep kernel method for lithofacies identification using conventional well logs. Pet. Sci. 20 (3), 1411–1428. https://doi.org/10.1016/j.petsci.2022.11.027.

Fang, X., Wang, Y., Zhang, Y., et al., 2024. Experimental investigation on identification of lithologies and layered rock interface based on drilling vibration response. Geoenergy Sci. Eng. 233, 212556. https://doi.org/10.1016/j.geoen.2023.212556.

Grana, D., Pirrone, M., Mukerji, T., 2012. Quantitative log interpretation and uncertainty propagation of petrophysical properties and facies classification from rock-physics modeling and formation evaluation analysis. Geophysics 77 (3), WA45–WA63. https://doi.org/10.1190/geo2011-0272.1.

Gu, Y., Bao, Z., Song, X., et al., 2019. Complex lithology prediction using probabilistic neural network improved by continuous restricted Boltzmann machine and particle swarm optimization. J. Petrol. Sci. Eng. 179, 966–978. https://doi.org/10.1016/j.petrol.2019.05.032.

Han, R., Wang, Z., Guo, Y., et al., 2023. Multi-label prediction method for lithology, lithofacies and fluid classes based on data augmentation by Cascade forest. Adv. Geo-Energy Res. 9 (1), 25–37. https://doi.org/10.46690/ager.2023.07.04.

He, H., Bai, Y., Garcia, E.A., et al., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, Hong Kong, China, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969.

He, M., Gu, H., Wan, H., 2020. Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir. J. Petrol. Sci. Eng. 194, 107498. https://doi.org/10.1016/j.petrol.2020.107498.

IEA, 2021. Net Zero by 2050: A Roadmap for the Global Energy Sector. International Energy Agency, Paris. https://www.iea.org/reports/net-zero-by-2050.

Ilkhchi, A.K., Rezaee, M., Ali Moallemi, S., 2006. A fuzzy logic approach for estimation of permeability and rock type from conventional well log data: An example from the Kangan reservoir in the Iran offshore gas Field. J. Geophys. Eng. 3 (4), 356–369. https://doi.org/10.1088/1742-2132/3/4/007.

Karimi, A.M., Sadeghnejad, S., Rezghi, M., 2021. Well-to-well correlation and identifying lithological boundaries by principal component analysis of well-logs. Comput. Geosci. 157, 104942. https://doi.org/10.1016/j.cageo.2021.104942.

Krivoshchekov, S., Kochnev, A., Kozyrev, N., et al., 2023. Rock typing approaches for effective complex carbonate reservoir characterization. Energies 16 (18), 6559. https://doi.org/10.3390/en16186559.

Lee, J., Lumley, D.E., 2023. Predicting shale mineralogical brittleness index from seismic and elastic property logs using interpretable deep learning. J. Petrol. Sci. Eng. 220, 111231. https://doi.org/10.1016/j.petrol.2022.111231.

Leite, V.R.C., Silva, P.M.C., Gattass, M., et al., 2013. Analysis of ensemble methods applied to lithology classification from well logs. In: 13th International Congress of the Brazilian Geophysical Society & EXPOGEF, Rio De Janeiro, Brazil, 26–29 August 2013. Brazilian Geophysical Society, pp. 949–952. https://doi.org/10.1190/sbgf2013-196.

Liu, M., Hu, S., Zhang, J., et al., 2023. Methods for identifying complex lithologies from log data based on machine learning. Unconv. Resour. 3, 20–29. https://doi.org/10.1016/j.uncres.2022.11.004.

Liu, X.Y., Zhou, L., Chen, X.H., et al., 2020. Lithofacies identification using support vector machine based on local deep multi-kernel learning. Pet. Sci. 17 (4), 954–966. https://doi.org/10.1007/s12182-020-00474-6.

Liu, Z., Ma, P., Wang, Y., et al., 2024a. KAN 2.0: Kolmogorov-Arnold networks meet science, arXiv:2408.10205. https://doi.org/10.48550/arXiv.2408.10205.

Liu, Z., Wang, Y., Vaidya, S., et al., 2024b. KAN: Kolmogorov-Arnold networks, arXiv:2404.19756. https://doi.org/10.48550/arXiv.2404.19756.

Lu, G., Zeng, L., Dong, S., et al., 2023. Lithology identification using graph neural network in continental shale oil reservoirs: A case study in Mahu Sag, Junggar Basin, western China. Mar. Petrol. Geol. 150, 106168. https://doi.org/10.1016/j.marpetgeo.2023.106168.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., et al. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., Red Hook, NY, pp. 4765–4774.

Ma, Y.Z., 2011. Lithofacies clustering using principal component analysis and neural network: applications to wireline logs. Math. Geosci. 43 (4), 401–419. https://doi.org/10.1007/s11004-011-9335-8.

Maiti, S., Tiwari, R.K., 2010. Neural network modeling and an uncertainty analysis in Bayesian framework: A case study from the KTB borehole site. J. Geophys. Res. 115 (B10), B10208. https://doi.org/10.1029/2010jb000864.

McCreery, E.B., Al-Mudhafar, W.J., 2017. Geostatistical classification of lithology using partitioning algorithms on well log data – A case study in Forest Hill oil field, East Texas Basin. In: 79th EAGE Conference and Exhibition 2017. European Association of Geoscientists & Engineers, Paris, France, pp. 1–5. https://doi.org/10.3997/2214-4609.201700905.

Mishra, A., Sharma, A., Patidar, A.K., 2022. Evaluation and development of a predictive model for geophysical well log data analysis and reservoir characterization: Machine learning applications to lithology prediction. Nat. Resour. Res. 31 (6), 3195–3222. https://doi.org/10.1007/s11053-022-10121-z.

Radziwill, N., 2019. Machine learning with R, third ed. (book Review): Brett lantz. 2019. Packt press, Birmingham, UK. 458 pages. Qual. Manag. J. 26 (4), 209. https://doi.org/10.1080/10686967.2019.1648086.

Rosid, M.S., Haikel, S., Haidar, M.W., 2019. Carbonate reservoir rock type classification using comparison of Naïve Bayes and Random Forest method in field "S" East Java. In: Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS 2018). AIP Publishing, Depok, Indonesia, 020019. https://doi.org/10.1063/1.5132446.

Saporetti, C.M., Duarte, G.R., Fonseca, T.L., et al., 2018. Extreme learning machine combined with a differential evolution algorithm for lithology identification. Revista de Informática Teórica e Aplicada 25 (4), 43–56. https://doi.org/10.22456/2175-2745.80702.

Song, L., Yin, X., Yin, L., 2023. Reservoir lithology identification based on improved adversarial learning. IEEE Geosci. Remote Sens. Lett. 20, 7503105. https://doi.org/10.1109/LGRS.2023.3281545.

Sun, Z., Jiang, B., Li, X., et al., 2020. A data-driven approach for lithology identification based on parameter-optimized ensemble learning. Energies 13 (15), 3903. https://doi.org/10.3390/en13153903.

Tolstaya, E., Shakirov, A., Mezghani, M., et al., 2023. Fast reservoir characterization with AI-based lithology prediction using drill cuttings images and noisy labels. J. Imaging 9 (7), 126. https://doi.org/10.3390/jimaging9070126.

Vale, K.M.O., Gorgônio, A.C., Da Luz E Gorgônio, F., et al., 2022. An efficient approach to select instances in self-training and co-training semi-supervised methods. IEEE Access 10, 7254–7276. https://doi.org/10.1109/ACCESS.2021.3138682.

van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440. https://doi.org/10.1007/s10994-019-05855-6.

Venieri, M., Harazim, D., Pedersen, P.K., et al., 2021. Vertical and lateral facies variability in organic-rich mudstones at the reservoir scale: A case study from the Devonian Duvernay Formation of Alberta, Canada. Mar. Petrol. Geol. 132, 105232. https://doi.org/10.1016/j.marpetgeo.2021.105232.

Wang, M., Fu, W., Hao, S., et al., 2016. Scalable semi-supervised learning by efficient anchor graph regularization. IEEE Trans. Knowl. Data Eng. 28 (7), 1864–1877. https://doi.org/10.1109/TKDE.2016.2535367.

Xie, Y., Zhu, C., Zhou, W., et al., 2018. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. J. Petrol. Sci. Eng. 160, 182–193. https://doi.org/10.1016/j.petrol.2017.10.028.

Yin, P., Qi, L., Zhu, Q., et al., 2018. Application of element logging to lithologic identification of key Horizons in Sichuan–Chongqing gas provinces. Nat. Gas Ind. B 5 (2), 132–138. https://doi.org/10.1016/j.ngib.2018.01.005.

Zhang, J., He, Y., Zhang, Y., et al., 2022a. Well-logging-based lithology classification using machine learning methods for high-quality reservoir identification: A case study of Baikouquan Formation in Mahu area of Junggar Basin, NW China. Energies 15 (10), 3675. https://doi.org/10.3390/en15103675.

Zhang, T., Guan, R., He, S., et al., 2026. A tutorial review on the fast pore-network extraction for modeling subsurface CO_2 sequestration. Energy. Sci. 2, 2630002. https://doi.org/10.1142/S2972379526300026.

Zhang, Y., Jiang, S., He, Z., et al., 2022b. Characteristics of heterogeneous diagenesis and modification to physical properties of upper Paleozoic tight gas reservoir in eastern Ordos Basin. J. Petrol. Sci. Eng. 208, 109243. https://doi.org/10.1016/j.petrol.2021.109243.

Zhu, L., Li, H., Yang, Z., et al., 2018. Intelligent logging lithological interpretation with convolution neural networks. Petrophysics 59 (6), 799–810. https://doi.org/10.30632/PJV59N6-2018A5.

Zhu, X., Goldberg, A.B., 2009. Introduction to Semi-supervised Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, 3. Morgan & Claypool Publishers, San Rafael, CA, pp. 1–130.

Zhu, X., Zhang, H., Zhu, R., et al., 2024. Classification with noisy labels through tree-based models and semi-supervised learning: A case study of lithology identification. Expert Syst. Appl. 240, 122506. https://doi.org/10.1016/j.eswa.2023.122506.

Zhuang, X., Wang, W., Su, Y., et al., 2025. Life-cycle prediction and optimization of sequestration performance in CO_2 mixture huff-n-puff development for tight hydrocarbon reservoirs. Appl. Energy 388, 125618. https://doi.org/10.1016/j.apenergy.2025.125618.