Advances in
Geo-Energy Research

Vol. 19, No. 1, p. 14-29, 2026

Original article

A machine learning framework for mineralogical composition assessment in unconventional formations

Batyrkhan Gainitdinov[1], Yury Meshalkin[1], Denis Orlov[1], Evgeny Chekhonin[1], Julia Zagranovskaya[2], Dmitri Koroteev[1], Yury Popov[1]

[1] *Center for Petroleum Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 121205, Russia*

[2] *Institute of Earth Sciences, Saint Petersburg State University, Saint-Petersburg 191123, Russia*

Keywords:
Machine learning
well logs
mineral composition prediction
unconventional reservoir
gradient boosting
thermal profiling
subsurface characterization

Cited as:
Gainitdinov, B., Meshalkin, Y., Orlov, D., Chekhonin, E., Zagranovskaya, J., Koroteev, D., Popov, Y. A machine learning framework for mineralogical composition assessment in unconventional formations. Advances in Geo-Energy Research, 2026, 19(1): 14-29.
https://doi.org/10.46690/ager.2026.01.02

Abstract:
Quantitative determination of mineralogy, through laboratory core studies and high-definition spectroscopic logging, is effective but underutilized due to cost and complexity. Unconventional formations present additional challenges, such as kerogen presence, heterogeneity, and anisotropy. This problem can be addressed by utilizing well logs and thermal profiling with specialized wrappers, such as multioutput regressor and regressor chain. Several machine learning models and strategies for combining well logs on multiscale data from an unconventional formation in West Siberia were tested to predict the mass and volumetric fractions of minerals obtained from the Litho Scanner. The gradient boosting regressor, wrapped in a regressor chain and combined with conventional well logs, demonstrated superior performance in predicting both mineral weight and volume fractions, effectively capturing the heterogeneity of the rock structure. A comparison between the machine learning-based model and the Litho Scanner showed an average discrepancy, measured by the root mean squared error for weight fraction, of 0.026 in the Bazhenov Formation. The relationship between certain minerals and the thermal properties of the rock was validated by assessing the importance of thermal core logging data for quartz and pyrite. Moreover, the volume fraction of the rock matrix, composed of total organic carbon and other minerals, was predicted more accurately by incorporating thermal core logging data. The mineral densities, required for obtaining mineral volumes, were determined by solving an optimization problem. Subsequently, a theoretical model was used to calculate thermal conductivity from the mineral volume fractions, revealing a significant similarity between the predicted and experimental values.

1. Introduction

Refining the mineralogy of rocks is essential for studying reservoir structure and gaining additional insights into rock properties. Mineral composition serves as a fundamental control on reservoir properties such as porosity, permeability, and water saturation, and provides critical context for interpreting well-logging measurements. Consequently, variations in mineral fractions directly influence the effectiveness of oil and gas field development strategies, as they reflect fundamental shifts in geological depositional environments and reservoir quality.

In unconventional oil and gas reservoirs, mineralogy assumes a particularly crucial role in determining the rock's response to geophysical measurements and its overall reservoir quality. A distinctive characteristic of these reservoirs is the presence of minerals with diverse physical and chemical structures, along with kerogen, which exhibits strong interrelationships with key thermophysical properties such as thermal conductivity and specific heat capacity (Clauser and

*Corresponding author.
E-mail address: Batyrkhan.Gainitdinov@skoltech.ru (B. Gainitdinov); Yury.Meshalkin@skoltech.ru (Y. Meshalkin); D.Orlov@skoltech.ru (D. Orlov); E.Chekhonin@skoltech.ru (E. Chekhonin); julia.zagranovskaya@gmail.com (J. Zagranovskaya).

Yandy Scientific Press

2207-9963 © The Author(s) 2025.
Received September 27, 2025; revised November 10, 2025; accepted December 6, 2025; available online November 10, 2025.

Huenges, 1995). These properties are vital for understanding thermal maturity and hydrocarbon generation potential. When coupled with the inherent heterogeneity of unconventional reservoirs, they significantly impact productivity and the efficiency of enhanced recovery techniques, particularly thermal methods.

The geological description of a reservoir fundamentally relies on the incorporation of log data (Serra, 1983). However, well log analysis presents substantial challenges in unconventional formations due to pronounced heterogeneity and anisotropy, especially in organic-rich source rocks. While high-definition spectroscopy tools such as Schlumberger's Litho Scanner enable quantitative mineral content measurement in boreholes, their application remains limited due to considerable operational costs. Alternatively, laboratory techniques on core samples, including X-ray diffraction (XRD) (Tucker, 1988), X-ray fluorescence (XRF) (Amosova et al., 2015), atomic emission spectroscopy (Tóth et al., 2017), thermal analysis (Barshad, 1965), and factor analysis (Kozlov and Fomina, 2018) provide detailed mineralogical information. Nevertheless, despite their accuracy, these methods are inherently discrete, failing to provide continuous mineralogy coverage along the wellbore, and represent relatively expensive procedures.

In response to these limitations, machine learning (ML) techniques have emerged as powerful tools for enhancing the interpretation of logging data, offering the potential to reduce exploration costs while maintaining robust accuracy (Martin et al., 2021; Kumar et al., 2022). Consequently, well logs are increasingly employed for rock type determination (Meshalkin et al., 2020) and mineralogy identification, which may target either quantitative weight fractions or qualitative mineral descriptions. The selection of optimal input features, however, remains an active area of investigation. Although a common methodology utilizes XRF data as inputs with XRD-derived mineralogy as targets (Rodriguez-Galiano et al., 2015; Kodikara et al., 2024; Yang et al., 2024), this approach faces limitations for continuous formation characterization, particularly in non-coring intervals. Accordingly, this research focuses on well-logging curves, which mitigate scale discrepancy issues compared to XRD and offer particular promise in unconventional formations (Barham and Zainal Abidin, 2023; Khan and Kirmani, 2024), especially when complemented by advanced mineralogical data from tools like the Litho Scanner.

Mineralogy detection maintains critical importance for analyzing low-permeability unconventional reservoirs (Cui et al., 2022), where identifying complex shale formations is complicated by inherent heterogeneity. The presence of minerals such as kerogen and pyrite introduces additional complexities in volume estimation due to their distinctive compositions. While previous research has emphasized the significance of determining the rock matrix in unconventional reservoirs (Cui et al., 2022; Hu et al., 2023; Yang et al., 2024), and although Temnikova et al. (2022) developed computational approaches for the Bazhenov Formation using core analysis, a significant research gap persists. Specifically, no prior studies have integrated well-logging with thermal data for mineralogy prediction in the Bazhenov Formation, thereby establishing the novelty of the present investigation.

The performance of mineral prediction models exhibits strong dependence on algorithm selection and hyperparameter configuration, with accurate prediction of certain minerals remaining particularly challenging (Nawal et al., 2023). While traditional regression methods often struggle to capture nonlinear relationships between rock properties and well logs in low-porosity shales, more sophisticated approaches including ensemble methods and artificial neural networks have demonstrated increasing success (Craddock et al., 2021; Zhou et al., 2021). Notably, Kim et al. (2020) established the effectiveness of ML techniques for lithology and mineralogy determination, with boosting methods outperforming traditional approaches due to superior nonlinear modeling capabilities. Despite these advances, consensus regarding the optimal algorithm and evaluation metrics remains elusive. Comparative studies reveal varied outcomes: Park et al. (2021) identified random forests as most effective for gas-hydrate-rich sediments, while Nawal et al. (2023) found Elastic Net superior for clay/carbonate prediction and neural networks optimal for quartz. Similarly, Barham and Zainal Abidin (2023) demonstrated ANN advantages over classical ML algorithms, while different evaluation metrics - such as root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2) - have been employed across studies Cui et al. (2022) and Hu et al. (2023). This methodological variability underscores the necessity of context-specific algorithm and metric selection based on regional characteristics, input features, and target minerals.

Furthermore, accurate mineral prediction necessitates appropriate feature selection, with Kodikara et al. (2024) emphasizing the critical importance of correlation analysis, which exhibits significant regional variation. Prediction accuracy depends substantially on input feature selection (Nawal et al., 2023), while proper data preprocessing and hyperparameter tuning remain equally crucial for preventing overfitting and enhancing model performance. Particularly relevant to this study, rock-forming minerals in unconventional formations display substantial variations in thermal properties (Hu et al., 2023). Quartz, characterized by high thermal conductivity, contrasts markedly with lower-conductivity minerals like clays and organic matter, with these thermal disparities influencing key reservoir management decisions (Clauser and Huenges, 1995). Nevertheless, direct connections between thermal data and mineralogy remain inadequately explored in recent literature.

To address these research gaps, this study aims to develop an effective framework for predicting mineralogical composition across depth intervals in unconventional formations, incorporating organic-rich source rocks through machine learning approaches. The investigation explores the integration of diverse well logs and thermal profiling data for accurate mineral determination, examines optimal combinations of conventional well logs with appropriate model selection and tuning, and demonstrates the derived mineralogy's contribution to thermal conductivity determination. The outcomes provide significant implications for unconventional resource assessment, contributing to advanced tool development in this

Table 1. Mean (Std) of the mineral weight fractions for each well based on Litho Scanner measurements.

Well No.	Mineral					
	Sid	Dol	Clc	Pyr	Cla	QFM
1	0.01	0.048	0.048	0.032	0.477	0.383
	(0.06)	(0.066)	(0.078)	(0.033)	(0.201)	(0.209)
2	0.0	0.037	0.035	0.034	0.398	0.493
	(0.0)	(0.062)	(0.068)	(0.04)	(0.127)	(0.132)
3	0.003	0.038	0.03	0.024	0.375	0.514
	(0.015)	(0.042)	(0.06)	(0.031)	(0.121)	(0.124)

domain.

The main contributions of this article are as follows:

- A novel data-driven workflow for predicting multi-mineral compositions in unconventional reservoirs through integration of conventional well logs and high-resolution thermal profiling data.
- First application of this workflow to the Bazhenov Formation, demonstrating that machine learning models (specifically Gradient Boosting Regressor with Regressor Chain) accurately predict mineral mass and volume fractions (RMSE = 0.026) from logging data.
- Comprehensive evaluation of classical tabular learners under identical blocked cross-validation conditions, confirming superior robustness and interpretability of GBR + RegressorChain.
- Implementation of a physics-based validation step linking predicted mineral fractions to thermal conductivity through theoretical compositional modeling with independent calibration on hold-out zones.
- Demonstrated enhancement of prediction stability through thermal profiling integration, reducing volume fraction RMSE from 0.046 to 0.039, confirming thermal sensitivity to mineralogical heterogeneity.

2. Methods

2.1 Data

The study area of the oil field is located in Western Siberia (Russia) and is composed of a thick layer of terrigenous sedimentary cover deposits from the Meso-Cenozoic age, overlying rocks of the Pre-Jurassic complex. The wells penetrate the Achimov, Bazhenov, Abalak, and Tyumen Formations, which consist of argillites, silicites (low-carbonaceous/carbonaceous, clayey), carbonate and siliceous-carbonate low-carbonaceous/carbonaceous rocks, high-carbonaceous rocks (clayey-siliceous, clayey-siliceous-carbonate), siltstones, coals, sandstones with carbonate cement, as well as thin laminations of sandstones, siltstones, and argillites (Gavrilov et al., 2015; Chekhonin et al., 2021; Postnikova et al., 2021). The Bazhenov formation rocks are organic-rich, with a TOC content of up to 24%. Details on mineral composition are given in Table 1.

For the experiments conducted in this research, data from three wells in the oil field were utilized (Table 2). The dataset includes logs obtained from well-logging methods, thermal data acquired by contactless optical scanning techniques (Popov et al., 2016), and mineral composition along with total organic carbon content obtained from the Litho Scanner tool. Conventional well logs, along with averaged thermal profiling, the anisotropy coefficient (K), and the rock matrix structure, are presented in Fig. 1.

In addition to the basic set of well logs, other logging measurements were included as input features for the prediction of mineralogy, as detailed in Table 3. The weight content of clay, coal (COA), calcite (CLC), dolomite (DOL), pyrite (PYR), quartz-feldspar-mica (QFM), and siderite (SID) which are the outputs of the ML-based model - was measured by neutron-induced gamma-ray spectrometry logging. The total weight fractions of the mentioned minerals sum to one.

To enhance the quality and standardization of well-logging data, a dedicated preprocessing pipeline for LAS files was implemented. The workflow includes:

- Mnemonic unification: Log identifiers from different data source were standardized using an internal dictionary to ensure consistent feature names across wells.
- Data cleaning: Invalid missing-value indicators were replaced with NaN; physical outliers beyond predefined valid ranges were removed; and long intervals with constant values (indicative of sensor artifacts) were filtered out.
- Unit standardization: All parameters were converted into a unified measurement system - e.g., borehole diameter to meters, gamma-ray readings to μR/h (adjusted for detector type), neutron porosity to standardized porosity units.
- Depth alignment: All logs were resampled onto a uniform 0.1 m grid to match the vertical sampling of mineralogical data. The high-resolution thermal profiles (mm scale) were averaged to this 10 cm step to ensure proper depth alignment and prevent overfitting to local fluctuations.
- Normalization: Gamma-ray and other skewed features were transformed using a log-normal distribution fit, followed by standard scaling within a $\pm 2\sigma$ window.

Missing values were handled via hybrid imputation: Isolated gaps shorter than 0.3 m were linearly interpolated, whereas longer gaps were left as NaN and excluded from the training blocks to prevent artificial smoothing. All features were standardized using StandardScaler, with normalization parameters computed on training wells only to avoid data leakage during blocked cross-validation. All preprocessing steps were implemented as modular functions, allowing selective application based on the quality of the input data. The final cleaned and aligned dataset was used for model training and subsequent blocked cross-validation.

2.2 Method and Theory

The prediction of mineral composition represents a multi-scale output, and this specific type of regression problem is referred to as multioutput regression. Thus, the application of

Table 2. Experimental conditions.

Category	Parameter	Notation	Well-logging method	Uncertainty
Well logs	Gamma-Ray	GR (API)	Gamma-ray spectrometry	±2%
	Bulk density	RHOZ ($\mathrm{g/cm^3}$)	Three-detector lithology density	±0.01 $\mathrm{g/cm^3}$
	Electrical resistivity	RXOZ ($\mathrm{Ohm \cdot m}$)	Array induction tool	±2%
	Thermal neutron porosity	TNPH (c.u.)	Compensated neutron logging	±6%
	Photo electric factor	PEFZ (b/e)	Photoelectric factor tool	±0.8 b/e
	Caliper	HCAL (mm)	Borehole size	±2.5 mm
	Compressional slowness	DTCO (ms/m)	Compressional slowness	±2%
	Shear slowness	DTSM (ms/m)	Shear slowness	±2%
	Bulk modulus	BMK (GPa)	Calculated via sonic and density logs	±5%
Thermal	Thermal conductivity $\parallel$	$\lambda_{\parallel}$ ($\mathrm{W/(m \cdot K)}$)	Optical scanning technique	±1.5%
	Thermal conductivity $\perp$	$\lambda_{\perp}$ ($\mathrm{W/(m \cdot K)}$)	Optical scanning technique	±1.5%
	Thermal diffusivity	a ($\mathrm{m^2/s}$)	Optical scanning technique	±2.0%
	Volumetric heat capacity	C ($\mathrm{J/(m^3 \cdot K)}$)	Calculated as $\lambda_{\parallel}/a$	±2.0%
	Thermal anisotropy	K (-)	Calculated as $\lambda_{\parallel}/\lambda_{\perp}$	
TOC	Total organic carbon	TOC (%)	Neutron-gamma ray spectrometry corrected to pyrolysis (LithoScanner tool)	±2%
Mineralogy	Coal	COA (c.u.)	Neutron-gamma ray spectrometry (LithoScanner tool)	±2%
	Calcite	CLC (c.u.)		
	Dolomite	DOL (c.u.)		
	Pyrite	PYR (c.u.)		
	Quartz-Feldspar-Mica	QFM (c.u.)		
	Siderite	SID (c.u.)		

wrapper methods serves as a workaround for models to predict multiple interdependent targets. These wrappers generally maintain the connections between the outputs, which is crucial since mineral fractions are dependent on each other. The Regressor Chain wrapper creates a linear sequence of models, where the output of each initial model serves as an input for the subsequent one. This chaining allows each model's prediction to influence the next in a sequential manner, thereby integrating a meaningful dependency structure to efficiently capture and utilize relationships among the structured multiple outputs. In contrast, a multioutput wrapper, such as the MultiOutput Regressor, treats outputs as independent and predicts each target variable separately. Some machine learning models natively support multiscale outputs (Breiman, 2001; Hastie et al., 2009; Hall et al., 2008), including K-Neighbors Regressor and Random Forest.

In this study, we have developed a comprehensive methodology for predicting several outputs. This methodology encompasses data preprocessing and feature combination, model selection and hyperparameter tuning, an evaluation of each algorithm's ease of application, and the use of wrapper techniques for models that do not natively support multiscale predictions (namely, Gradient Boosting Regressor, LightGBM, CatBoost, XGBoost.). The models listed above that are inherently capable of handling multiple outputs were also implemented. This overall approach contrasts with state-of-the-art methods commonly applied for mineralogy prediction from well-logging data. Moreover, thermal data was incorporated as an input for several modeling strategies. In this research, RMSE and MAE were used as evaluation metrics for predicting the minerals' weight fractions, as defined in Chai and Draxler (2014) (Eqs. (1) and 2). The metrics were calculated using the mineral weight fractions from Litho Scanner measurements (w_{ij}) and the corresponding model predictions ($\hat{w}_{ij}$), considering various mineral components (indexed by "j") across different depth intervals (indexed by "i").

$$\mathrm{RMSE} = \sqrt{\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(w_{ij} - \hat{w}_{ij}\right)^2} \tag{1}$$

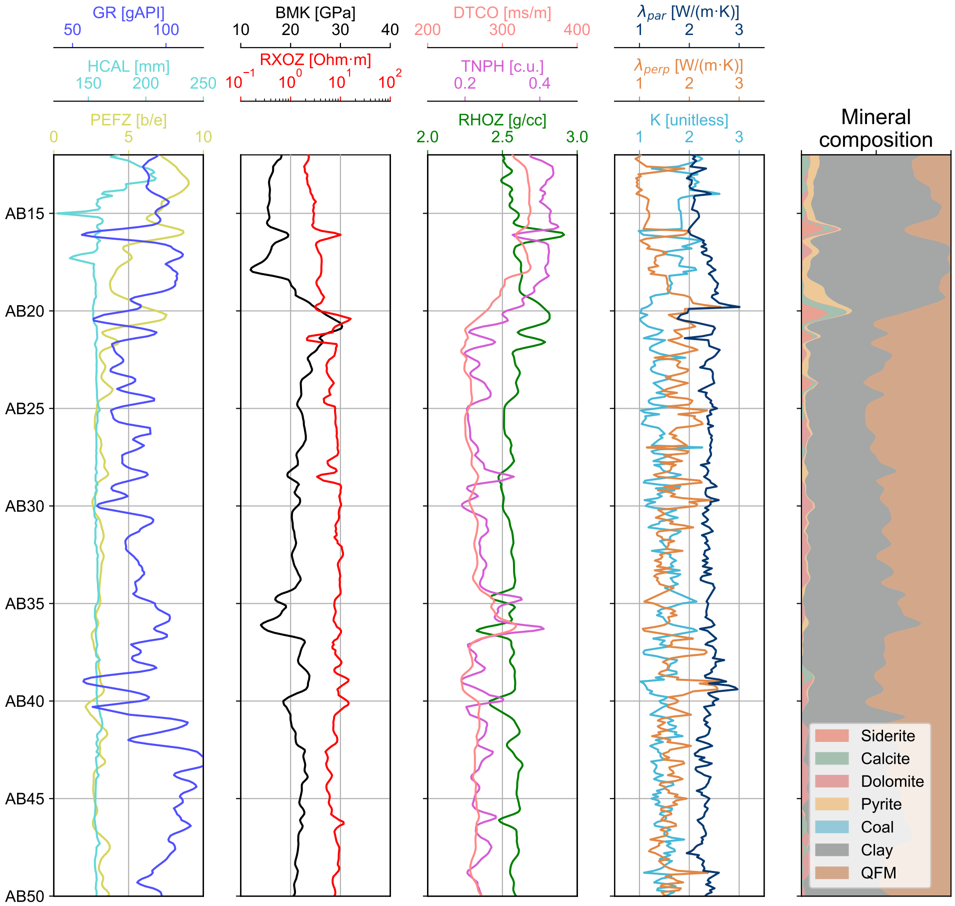


Fig. 1. Example of well logging curves and mineralogical composition on 30 m interval.

$$\mathrm{MAE} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|w_{ij} - \hat{w}_{ij}| \tag{2}$$

RMSE and MAE are widely used in regression tasks for evaluating model accuracy. RMSE imposes a higher penalty on large errors due to its quadratic nature, while MAE provides a linear and more straightforward average error. Together, these metrics offer comprehensive insights into model performance, capturing both the magnitude of typical errors and the impact of significant deviations. This dual assessment is particularly appropriate for predicting continuous variables such as mineralogical composition, where both overall accuracy and extreme deviations are critical.

The study employed several validation strategies, including training on one well and predicting on others, as well as training on two wells and validating on a third. For each strategy outlined in Section 3.2.1, a systematic input data selection process was conducted. This included evaluating model performance with and without the inclusion of thermal data in addition to conventional well logs.

The output of the Regressor Chain and MultiOutput Regressor models represent predicted mineral mass fractions. Since mineralogical compositions are by definition compositional (non-negative and summing to one), several approaches were tested to ensure physically consistent predictions:

- Additive log-ratio (ALR) transformation.
- Softmax/logit mapping.
- Simple post-scaling normalization.
- Euclidean simplex projection following the algorithm of (Wang and Carreira-Perpinan, 2013).

Models that natively support multi-output regression are capable of predicting mineral fractions in a manner that constrains their sum to unity. This inherent capability represents an advantage over wrapper-based approaches, which may not automatically preserve such physical constraints in the predictions.

The workflow for mineral prediction and subsequent utilization of the model output for thermal conductivity (TC) calculation is presented in Fig. 2.

Fig. 2(a) corresponds to the best strategy selection for mineral weight fraction prediction in terms of the model and data combination. Among the available thermal profiling and well logging data, the best parameters are chosen (Section 4.1). To assess the best model several strong tabular baselines were implemented under the same experimental setup. These included Random Forest, LightGBM, CatBoost, XGBoost, and k-Nearest Neighbors (KNN) regressors. The best model and set of well logs are further used in Fig. 2(b) for the prediction of mineral volume fractions, which are required for the calculation of TC via a theoretical model. The thermal properties of the rock are highly dependent on the presence of kerogen. The conversion of mass to volume fractions is com-

Table 3. Additional well logging data participating in mineralogy determination.

Type	Model
HCGR	Computed gamma-ray (API)
DSTST	Stoneley slowness (ms/m)
VPVS	Compressional to shear ratio (-)
DTSH	Stoneley shear slowness (ms/m)
FSH_P1NO	Fast shear azimuth (deg)
AT 90	Resistivity at 90 inch radial midpoint (Ohm·m)
AT 60	Resistivity at 60 inch radial midpoint (Ohm·m)
AT 20	Resistivity at 20 inch radial midpoint (Ohm·m)
HSGR	Total gamma-ray (API)
PR	Poisson ratio (-)
MRP_LHR	Magnetic resonance porosity (-)
BFV_LHR	Bound fluid volume for lower high-resolution antenna (-)
MAXXENE_OVERALL	Maximum cross-dipole energy (-)
MINXENE_OVERALL	Minimum cross-dipole energy (-)

pleted using bulk density, TOC, parameter S, and mineral densities, as described in Section 3.2.2. Thus, at this stage, it is verified whether the addition of thermal profiling to well logging data improves the results. In Fig. 2(c), the application of the theoretical model is considered (Section 3.2.3) for the calculation of TC based on predicted volume fractions and input data from Fig. 2(b). The result is finally compared to the parallel component of the TC tensor.

2.2.1 Strategies for mineral weight fraction prediction

The development of strategies was based on available data, comprising well logging, thermal profiling data, and mineral weight content, including kerogen. The first three strategies were designed for cases where the model was trained and tested on a single well, i.e., a case of data scarcity; one can see the considered strategies below:

- It is composed of various types of well logs, including conventional logs, Array-Sonic logs, and advanced logging techniques such as Nuclear Magnetic Resonance profiles and Electrical Microimager. Additionally, petrophysical property measurements, along with mineral weight content, were utilized for both training and testing purposes;
- The training and testing of the model were performed using exclusively conventional well logging data. The selection of these logs was based on importance measurements, which determined their significant contribution to the predictive model (see Section 2 for details);
- Building upon strategy 2 by incorporating thermal data into the previous scenario.

The remaining strategies use two wells for training and one for testing:

- Utilizing two wells for training and one well for testing, with only conventional well logging data as input;
- Improving upon strategy 4 by compositional correction via Euclidean projection onto the simplex;
- Extending strategy 5 by adding thermal data to the analysis.

See the results of each strategy prediction in Section 4.

2.2.2 Model validation and data splitting

To ensure a realistic depth-wise validation and avoid optimistic bias due to autocorrelation along depth, we implemented blocked cross-validation (blocked CV) within each well.

For strategies 1-3 (see Section 2.2.1) (where a single well was used for both training and testing), the well depth was divided into non-overlapping blocks of 10-20 m, separated by gap zones of 4 m to eliminate correlation between neighboring intervals (Fig. 3). At each CV iteration, one block was used for validation, while the remaining blocks served as training. Additionally, several small depth intervals were reserved as "physics" regions, used exclusively for calibrating mineral densities and thermal conductivities (see Sections 2.2.3 and 2.2.4).

For strategies 4-6 (two training wells and one held-out test well), the blocked CV was applied only within the training wells to tune hyperparameters of the model (Fig. 4). The final model was then retrained on all training blocks (excluding “physics” intervals) and evaluated on the held-out well.

This design guarantees that no depth intervals adjacent in depth or belonging to the same physical calibration region contribute simultaneously to training and validation, ensuring a clean separation between data used for model fitting, hyperparameter tuning, and physical parameter calibration.

2.2.3 Mineral density estimation

Mineral densities are needed to convert mineral masses into volumes. Often, the lack of this information leads to the search for methods for estimating densities. In this paper, a mathematical optimization method was proposed (Alekseev and Gavrilov, 2019). To a reasonable approximation, the fluid-saturated rock of an unconventional reservoir consists of the fluid itself, a matrix of minerals, and organic matter, mainly kerogen. Therefore, bulk density (ρ_b) can be represented as in Eq. (3). The provided data includes fluid density (ρ_f = 0.6 g/cm^3), porosity (ϕ), kerogen density (ρ_k = 1.4 g/cm^3) (Dang et al., 2016), volume fraction of kerogen (c_k), and mineral matrix density (ρ_m). The volume fraction of kerogen is calculated using Eq. (4), where TOC is the amount of organic carbon present in a source rock (expressed as a weight percent), used as a proxy for the total amount of organic matter (kerogen) present in the sediment, and parameter p is the mass

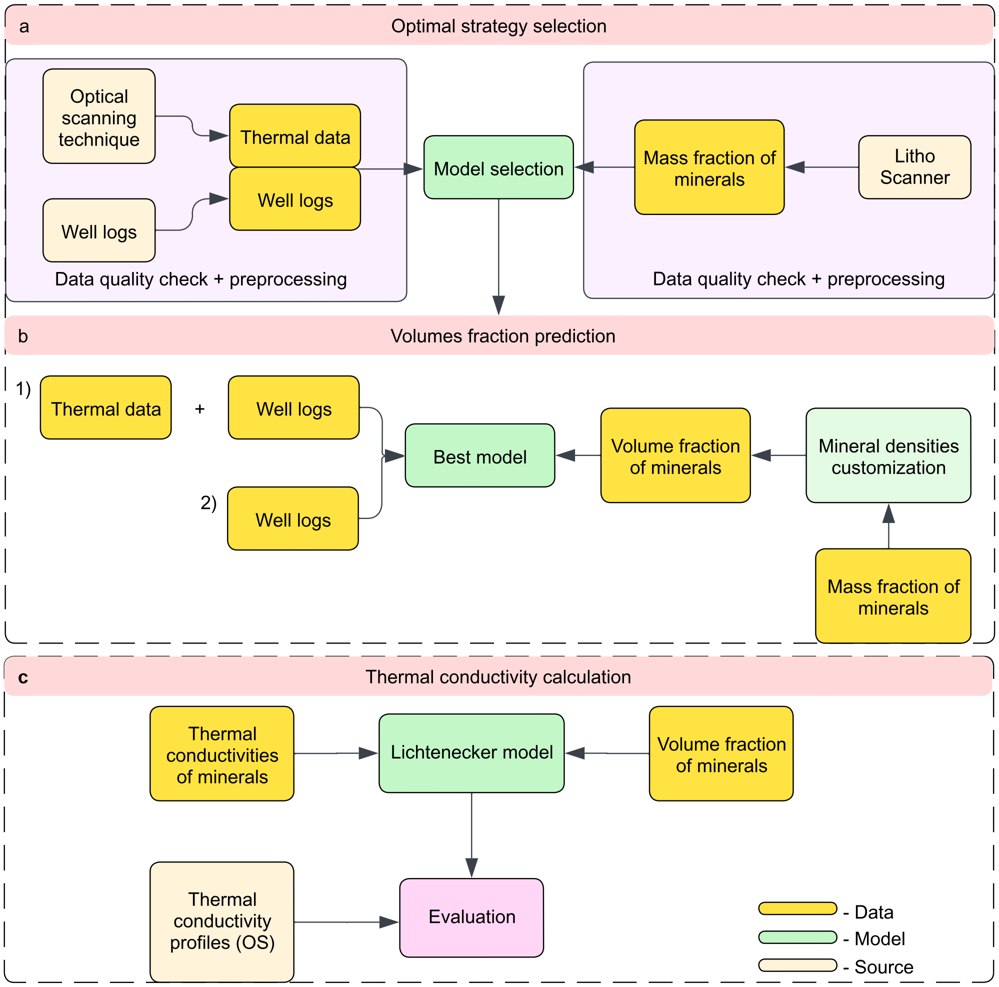

a
Optimal strategy selection
Optical scanning technique
Thermal data
Well logs
Well logs
Data quality check + preprocessing
Model selection
Mass fraction of minerals
Litho Scanner
Data quality check + preprocessing
b
Volumes fraction prediction
1)
Thermal data
+
Well logs
2)
Well logs
Best model
Volume fraction of minerals
Mineral densities customization
Mass fraction of minerals
c
Thermal conductivity calculation
Thermal conductivities of minerals
Lichtenecker model
Volume fraction of minerals
Thermal conductivity profiles (OS)
Evaluation
- Data
- Model
- Source

Fig. 2. Work pipeline scheme. (a) The best strategy selection for mineral weight fraction prediction, (b) the best model and set of well logs for the prediction of mineral volume fractions and (c) the calculation of thermal conductivity based on predicted volume fractions and input data.

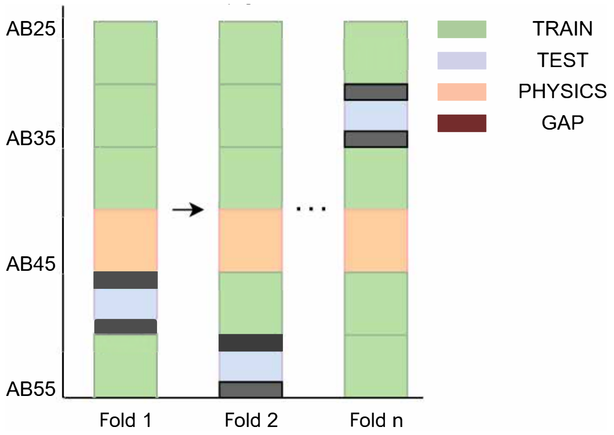


Fig. 3. Blocked CV scheme for 1-3 strategies.

content of carbon in hydrocarbons (Hantschel and Kauerauf, 2009) (p = 0.87). Based on Eqs. (3) and 4, an expression for the density of minerals has been obtained, Eq. (5):

$$\rho_b = \rho_f \phi + \rho_k c_k + \rho_m (1 - \phi - c_k) \tag{3}$$

$$c_k = \frac{\text{TOC}}{100p} \frac{\rho_b}{\rho_k} \tag{4}$$

$$\rho_m = \frac{\rho_b - \rho_f \phi - \rho_b \dfrac{\text{TOC}}{100p}}{1 - \phi - \dfrac{\text{TOC}}{100p} \dfrac{\rho_b}{\rho_k}} \tag{5}$$

The objective function for determining mineral densities (ρ_i, where i is the mineral index) and rock porosity by depth (j index) is presented in Eq. (6). To solve the optimization problem, the trust-region constrained algorithm (Trust-constr) (Conn et al., 2000) was applied within the following ranges of acceptable values: $\phi \in [0, 0.25]$, $\rho_{ker} \in [1.1, 1.3]$, $\rho_{cla} \in [2.5, 3]$, $\rho_{clc} \in [2.5, 2.7]$, $\rho_{dol} \in [2.6, 2.9]$, $\rho_{pyr} \in [4.5, 5.1]$, $\rho_{QFM} \in [2.2, 3.3]$, $\rho_{sid} \in [3.7, 3.9]$.

$$\min_{\rho_i, \phi} \sum_{j=1}^{n} \left(\sum_{i=1}^{k} \frac{w_{ij} \rho_{mj}}{\rho_i} - 1 \right)^2 \tag{6}$$

The optimization of mineral densities ρ_{mj} was performed exclusively on the designated hold-out intervals, which were

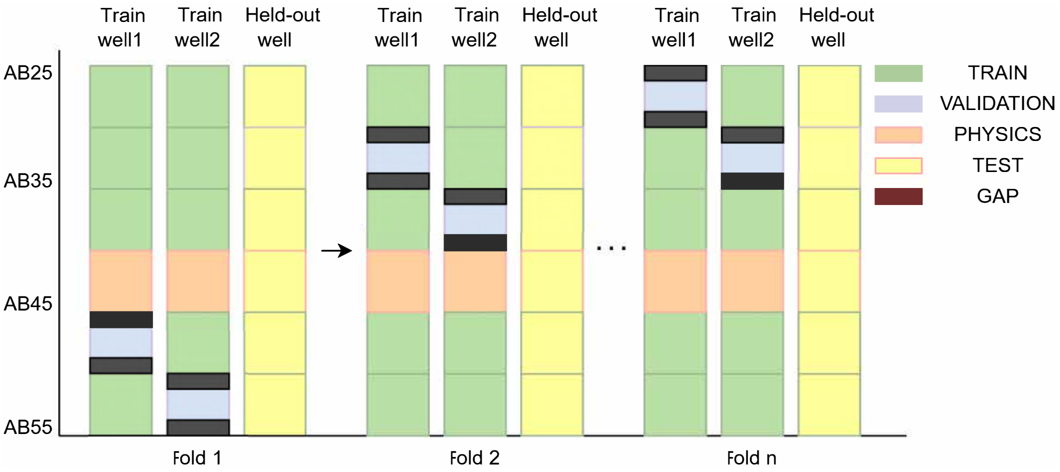


Fig. 4. Blocked CV scheme for 4-6 strategies.

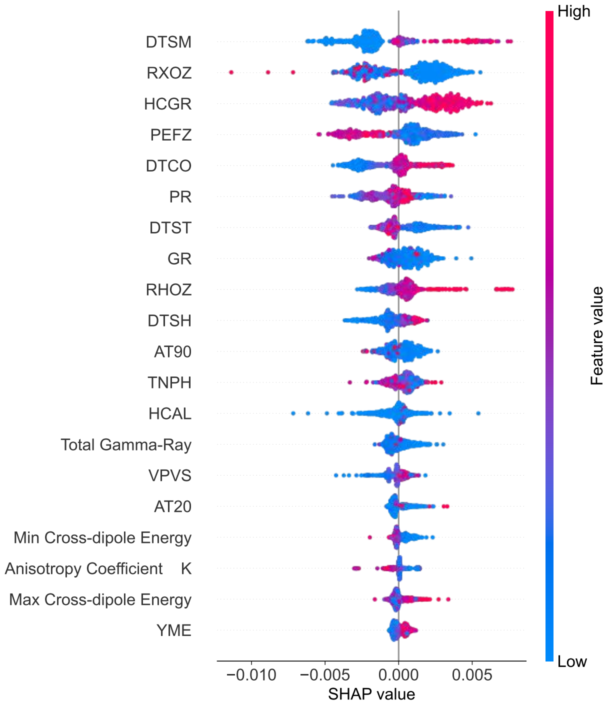


Fig. 5. SHAP summary plot showing global feature importance across all predicted minerals.

separated from both training and validation blocks (see Fig. 4). This ensures that the density calibration does not influence the ML model training or evaluation and thus maintains the independence of the physics-based verification step.

2.2.4 Thermal conductivity calculation

As mentioned earlier, to evaluate the accuracy of mineral volume fraction prediction, the study applied a theoretical model (Eq. (7)) commonly used for assessing thermal conductivity based on model predictions. Thus, the model's performance can be evaluated not only by checking RMSE and MAE but also by examining the results of a subtask utilizing the model outputs. To account for the difference in spatial resolution between the optical scanning instrument and well-logging tools, the initial 1-mm thermal profiling data was averaged using a 50-cm moving window. The average porosity value along the depth is below 4%, and as a result, it was ignored in the formula for thermal conductivity calculation:

$$\lambda_{eff} = \lambda_{ker}^{V_{ker}} \prod_{i=1}^{N} \lambda_i^{V_i} \tag{7}$$

where V_{ker} and V_i are the kerogen and i-th mineral component volume fractions, respectively; λ_{eff}, λ_{ker} and λ_i are the thermal conductivity (TC) of the effective rock, kerogen, and i-th component of the rock matrix, respectively. The TC of minerals is chosen to minimize the discrepancy between the experimental TC curve and the TC calculated from model predictions; these values are determined within predefined ranges: $C_{sid} \in [3.0, 3.1]$, $C_{dol} \in [4.9, 6.3]$, $C_{clc} \in [3.1, 3.6]$, $C_{pyr} \in [19.2, 41.4]$, $C_{cla} \in [1.2, 2.7]$, $C_{QFM} \in [2.1, 7.6]$ (Sass, 1965; Horai, 1971; Beck et al., 1978; Popov et al., 1987).

It is essential to note that the TC of the QFM mineral component varies between areas inside and outside the Bazhenov Formation. The optimal TC value was chosen based on these considerations. Moreover, we applied a curve smoothing approach using the Fast Fourier Transform technique (Brigham, 1988) to the experimental TC, the TC calculated from Litho Scanner measurements, and the model predictions. The smoothed curves exhibit enhanced clarity, reduced local variations, and better focus on trends. This enables more accurate analysis and interpretation of the data, providing valuable insights into the material's thermal properties.

3. Results

3.1 Optimal strategy selection

To assess the influence of different well logs on the prediction of mineral compositions, feature selection was performed on the full set of input data (Fig. 5), including both conventional and additional well logs (Tables 2 and 3). To improve interpretability and ensure that the ML model captures physically meaningful dependencies, SHAP analysis was performed on the GradientBoosting + RegressorChain model. Fig. 5 illustrates the relative importance of each fea-

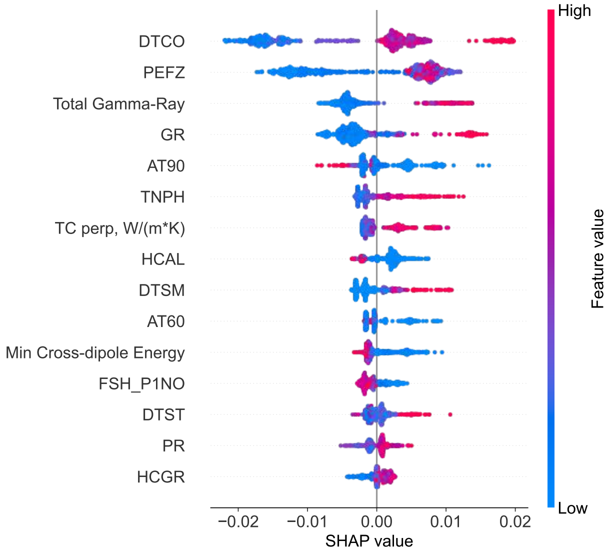


Fig. 6. SHAP summary plot for Pyrite illustrating key petrophysical controls on mineral variability.

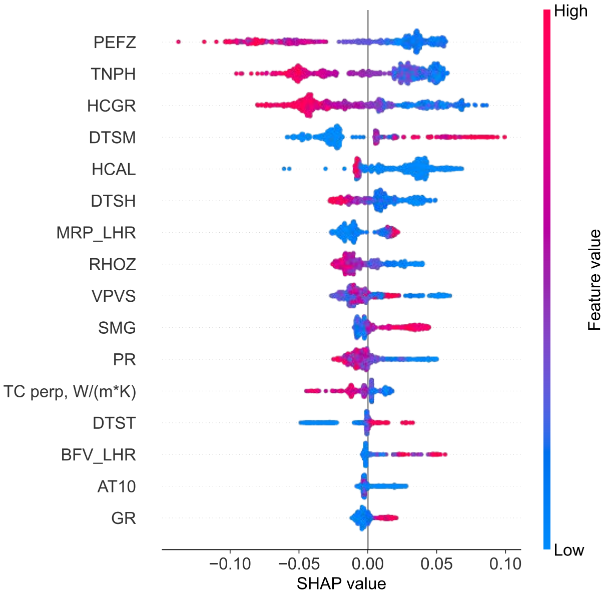


Fig. 7. SHAP summary plot for QFM illustrating key petrophysical controls on mineral variability.

ture in the model's output. The SHAP analysis confirms that total gamma-ray (HSGR), bulk density (RHOZ), and sonic velocity (DTCO) are the dominant features across most minerals, in line with known petrophysical trends. Moreover, all conventional well logging measurements demonstrate high SHAP values. Based on these findings and typical standard well measurements conducted in the Bazhenov Formation, the final set of well logs was selected (Table 2).

As discussed earlier, specific minerals are closely tied to the thermal properties of the rock. To validate this for quartz and pyrite, SHAP analysis revealed that the perpendicular component of thermal conductivity had a more significant influence on model predictions for these minerals than for others. The SHAP plot (Figs. 6 and 7) illustrates the dominant role of thermal properties in determining quartz and pyrite concentrations, reinforcing their importance in the study of complex carbonate reservoirs. Pyrite prediction is primarily driven by DTCO and PEFZ, whereas QFM is governed by HCGR, and neutron-density logs. Thermal and cross-dipole energy attributes show secondary influence, particularly at greater depths, suggesting their role in modulating anisotropy and siderite variability. These relationships demonstrate that the model captures physically meaningful dependencies rather than statistical artifacts.

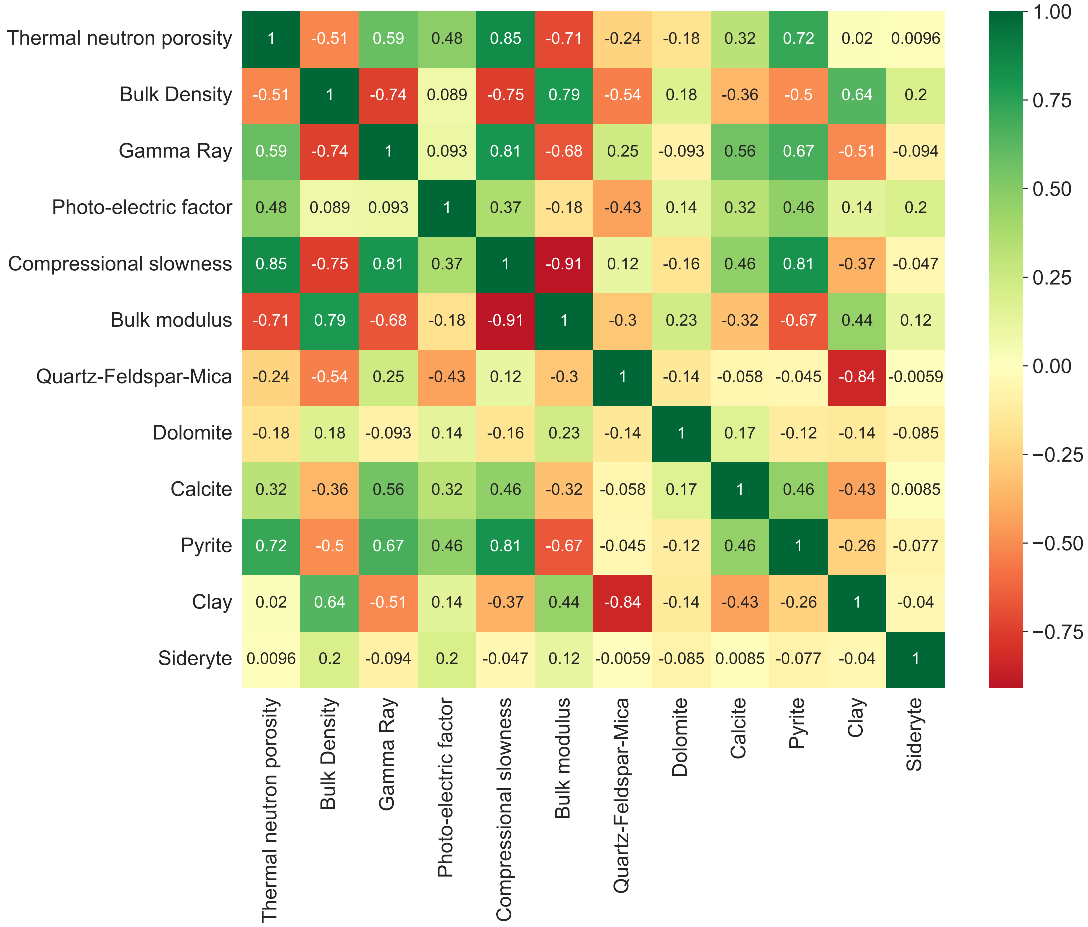
Based on the importance of selected features and the logs used by Kim et al. (2020), the parameters for heatmap construction were chosen to compare with their findings. The correlation largely aligns, with some exceptions. The heatmap (Fig. 8) shows a positive relationship between DTSO and TNPH, GR, calcite, and pyrite, while it exhibits a negative correlation with RHOZ, dolomite, and total clay.

This confirms that conventional well logs capture the trends of mineralogy variation and show a clear correlation between well logs and mineralogy content, as also indicated by Hu et al. (2023). Furthermore, Fig. 1 reveals that increasing QFM and decreasing clay content correlate with lower TNPH, DTCO, and PEFZ values and an increase in BMK. A rise in pyrite content reflects similar oscillations in the anisotropy coefficient, while variations in calcite and dolomite fractions are associated with RXOZ, PEFZ, and RHOZ.

The prediction results for the strategies (see Section 3.1) are presented in Table 4, which shows the model performance for each case. The metrics in the table represent the mean MAE and RMSE values for both blocked CV strategy and held-out test well for the weight fraction prediction of all six mineral classes.

The blocked CV procedure reduced the risk of overfitting to depth-contiguous intervals and provided more conservative RMSE and MAE estimates. Consistent trends were observed across strategies: Intra-well validation (strategies 1-3) produced lower apparent RMSE, while inter-well evaluation (strategies 4-6) yielded higher but more realistic macro-RMSE values, confirming that the model generalizes beyond individual well depth profiles.

To evaluate the influence of compositional enforcement on model performance, several transformations (ALR, softmax, normalization, and Euclidean projection) (see Section 2.2) were tested on the same training/validation setup for the 5th strategy. The ALR and softmax transforms did not improve accuracy and in some cases led to unstable training on the limited dataset. In contrast, the simple normalization and the Euclidean simplex projection improved the predictive accuracy while strictly enforcing compositional validity ($w_i \geq 0$, $\sum w_i = 1$, where w_i is the i^{th} mineral mass fraction). Therefore, in the final workflow the Euclidean simplex projection was adopted as the standard post-processing step applied to all model outputs. The prediction results of Strategy 5 are shown in Fig. 9. The compositional projection slightly sharpens relative proportions but does not alter overall mineral trends with depth, confirming that the base model already captures correct relative relationships, and the adjustment serves mainly to enforce formal compositional consistency. The predicted mineral composition is very close to the actual composition



	Thermal neutron porosity	Bulk Density	Gamma Ray	Photo-electric factor	Compressional slowness	Bulk modulus	Quartz-Feldspar-Mica	Dolomite	Calcite	Pyrite	Clay	Sideryte
Thermal neutron porosity	1	-0.51	0.59	0.48	0.85	-0.71	-0.24	-0.18	0.32	0.72	0.02	0.0096
Bulk Density	-0.51	1	-0.74	0.089	-0.75	0.79	-0.54	0.18	-0.36	-0.5	0.64	0.2
Gamma Ray	0.59	-0.74	1	0.093	0.81	-0.68	0.25	-0.093	0.56	0.67	-0.51	-0.094
Photo-electric factor	0.48	0.089	0.093	1	0.37	-0.18	-0.43	0.14	0.32	0.46	0.14	0.2
Compressional slowness	0.85	-0.75	0.81	0.37	1	-0.91	0.12	-0.16	0.46	0.81	-0.37	-0.047
Bulk modulus	-0.71	0.79	-0.68	-0.18	-0.91	1	-0.3	0.23	-0.32	-0.67	0.44	0.12
Quartz-Feldspar-Mica	-0.24	-0.54	0.25	-0.43	0.12	-0.3	1	-0.14	-0.058	-0.045	-0.84	-0.0059
Dolomite	-0.18	0.18	-0.093	0.14	-0.16	0.23	-0.14	1	0.17	-0.12	-0.14	-0.085
Calcite	0.32	-0.36	0.56	0.32	0.46	-0.32	-0.058	0.17	1	0.46	-0.43	0.0085
Pyrite	0.72	-0.5	0.67	0.46	0.81	-0.67	-0.045	-0.12	0.46	1	-0.26	-0.077
Clay	0.02	0.64	-0.51	0.14	-0.37	0.44	-0.84	-0.14	-0.43	-0.26	1	-0.04
Sideryte	0.0096	0.2	-0.094	0.2	-0.047	0.12	-0.0059	-0.085	0.0085	-0.077	-0.04	1

Fig. 8. Heatmap showing the correlation relationships between features and labels. Green and red indicate positive and negative correlation coefficients, respectively.

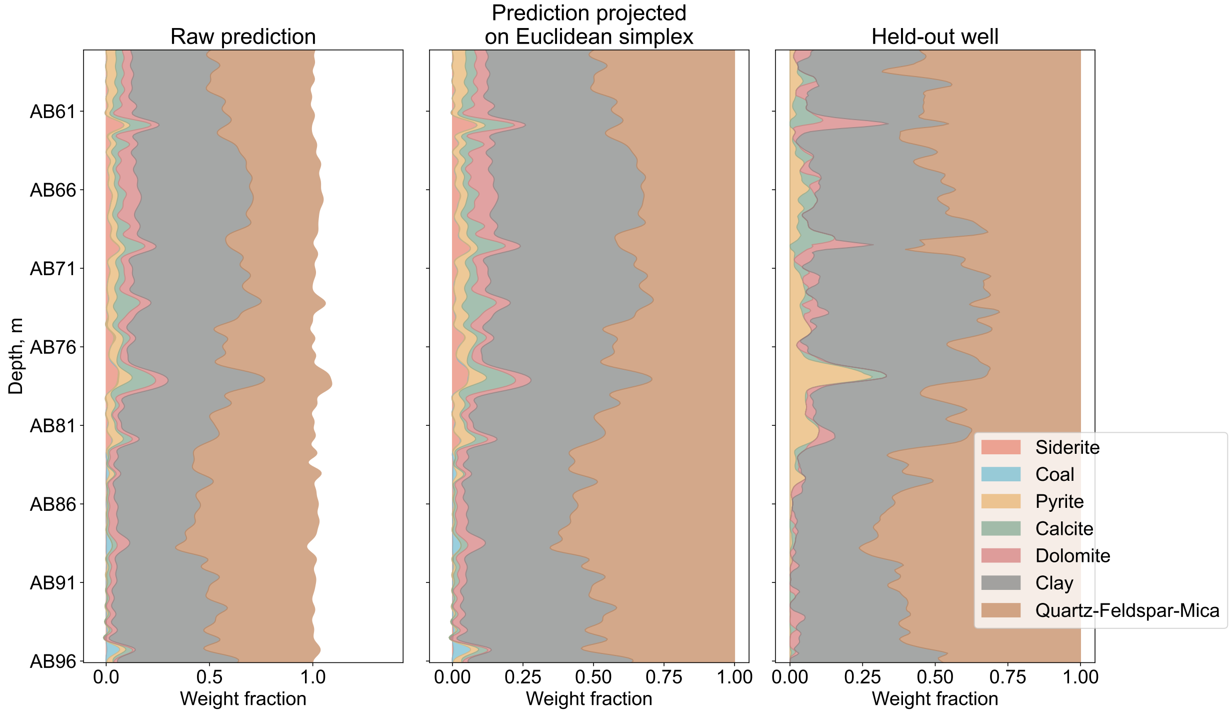


Fig. 9. The prediction of mineral weight fractions on 30 m depth interval for the 5th strategy in the 1st well.

Table 4. Comparison of model performance for weight fraction prediction across ablation strategies.

Strategy	Blocked CV MAE	Blocked CV RMSE	Held-out well MAE	Held-out well RMSE
1	0.0214	0.0297	/	/
2	0.0231	0.0327	/	/
3	0.0234	0.0312	/	/
4	0.0235	0.0371	0.028	0.038
5	0.0218	0.0333	0.026	0.035
6	0.0217	0.0322	0.025	0.034

Notes: Strategies 1-3 single train-test well, while strategies 4-6 use 2 train and 1 held-out test well. Strategy 5 (GBR. + RegressorChain) yields the lowest overall RMSE.

Table 5. Comparison of Basic algorithms for the 5th strategy.

Algorithm	Blocked CV RMSE	Blocked CV MAE	Held-out well RMSE	Held-out well MAE
GBR	0.0333	0.0218	0.035	0.026
LGBM	0.0344	0.0236	0.0338	0.0265
CB	0.032	0.0214	0.0357	0.0268
XGB	0.0331	0.0228	0.0357	0.025
RF	0.035	0.0241	0.036	0.0273
KNN	0.0363	0.0242	0.0423	0.0299

obtained via Litho Scanner. At certain depth intervals, the difference between the actual measurement and the prediction can be up to 20%, but the average deviation does not exceed 2%. The predicted values also accurately replicate the trend of mineral composition with depth, indicating sensitivity to rock type.

Addition of thermal data for prediction of mineral weight fraction as for the 6th strategy doesn't give significant improvements in comparison to previous strategy.

In the process of model selection, several machine learning algorithms were evaluated, including GradientBoostingRegressor (GBR), LightGBM (LGBM), CatBoost (CB), XGBoost, K-Nearest Neighbors (KNN) and Random Forest (RF). All models were trained following the 5th strategy (two training wells, one held-out test well) and validated using the same blocked cross-validation scheme within the training wells. The evaluation metrics were RMSE and MAE, computed both for blocked CV folds and for the held-out test well to ensure fair comparison (Table 5). Hyperparameters were tuned within conservative ranges using blocked CV to prevent overfitting to depth-wise correlations. The results indicate that all ensemble-based learners perform comparably in terms of blocked CV RMSE (≈0.033-0.035) and held-out well RMSE (≈0.033-0.036), while the Gradient Boosting Regressor (GBR) slightly outperforms other methods in both metrics. CatBoost and LightGBM show close performance but with slightly higher variance across validation folds, whereas K-nearest neighbor regression degrades on the held-out well due to its local, non-parametric nature.

Table 6. Comparison of mineral weight fraction predictions for GBR (Strategy 5) in terms of R^2 and RMSE.

Mineral	R^2	RMSE
Cla	0.959	0.043
Clc	0.956	0.019
Dol	0.792	0.023
Pyr	0.958	0.008
QFM	0.946	0.044
Sid	0.774	0.009

Given the small dataset and strong inter-feature correlations typical of well-log data, the GBR-RegressorChain combination offers a robust trade-off between accuracy, interpretability, and stability. Unlike CatBoost or LightGBM, which internally encode categorical structure, GBR allows explicit feature importance and SHAP-based interpretability analysis, facilitating physical interpretation of input contributions. Therefore, GBR + RegressorChain was retained as the primary model for all subsequent compositional and thermal-conductivity analyses.

As mentioned above, each output label was predicted separately. The prediction quality for each mineral was evaluated using R^2 and RMSE (see Table 6). Clay and QFM show the highest RMSE values for weight fraction prediction.

For the implementation of the algorithm, Python and Scikit-learn's GradientBoostingRegressor were used, while RegressorChain was provided by the Multioutput module of Scikit-learn. The model, trained using a GradientBoostingRegressor wrapped in a RegressorChain, was run on a dataset with 51 features (including thermal data) for approximately 400 m depth interval with a measurement resolution of 10 cm and with 6 mineralogical output targets, and tested on a 200 m depth interval. For the GradientBoostingRegressor wrapped in a RegressorChain, key hyperparameters include the learning rate, max depth, max features, and the number of estimators. To validate the robustness of the machine learning model and reduce the risk of overfitting, 5-fold cross-validation was implemented using GridSearchCV. The resulting optimal values were: Learning rate of 0.01, max depth of 5, sqrt for max features, and 50 estimators. The training process utilized approximately 120 MB of memory and was CPU-bound.

3.2 Thermal conductivity calculation

Prior to the thermal conductivity assessment, the model was retrained using volumetric mineral fractions as target variables. These fractions were obtained by converting ground-truth weight fractions from the Litho Scanner using mineral densities calibrated exclusively on the hold-out intervals, which were separated from both training and validation blocks. This ensures that the conversion parameters were tuned inde-

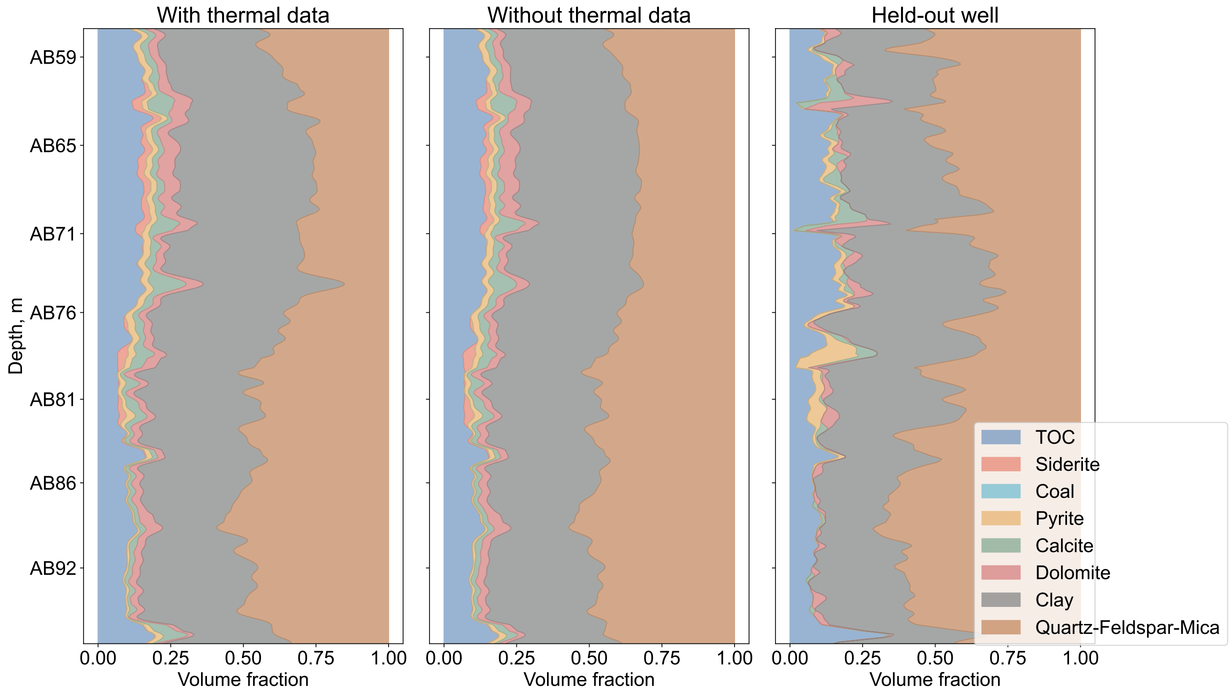


Fig. 10. The prediction of mineral volume fractions with and without thermal data for the held-out well.

Table 7. Customized densities and TC of minerals.

Mineral	Optimal density (g/cm^3)	Density ranges (-)	Thermal conductivity (W/(m·K))	Thermal conductivity ranges (-)
Sid	3.8	(3.7, 3.9)	3.1	(3.0, 3.1)
Dol	2.8	(2.6, 2.9)	5.3	(4.9, 6.3)
Clc	2.6	(2.5, 2.7)	3.1	(3.1, 3.6)
Pyr	4.8	(4.5, 5.1)	24.8	(19.2, 41.4)
Cla	2.6	(2.5, 3.0)	1.5	(1.2, 2.7)
QFM	2.5	(2.2, 3.3)	3.8; 3.3*	(2.1, 7.6*)

Notes: * Upper limit for TC of QFM minerals within the Bazhenov formation (3.3 W/(m·K)) is less than outside it (7.6 W/(m·K)) because lack of high-conductive quartz in the Bazhenov formation.

Table 8. RMSE for experimental TC vs TC calculated based on model volume fraction prediction and TC of minerals calibrated on a hold-out zones/test well.

Formation	Calibration on test well	Calibration on hold-out zones
Bazhenov	0.052	0.079
Abalak	0.170	0.132
Tyumen	0.243	0.154

pendently of the ML model, maintaining the integrity of the physics-based validation. Consequently, the model predicted volumetric mineral fractions directly rather than through post hoc conversion from weight fractions, providing a consistent foundation for the subsequent thermal conductivity analysis.

As a result, the initial focus was on identifying the optimal approach for predicting these volume fractions. Given that the presence of TOC plays a crucial role in defining the thermal properties of the rock, the reciprocal relationship was also considered. Therefore, an investigation was undertaken to assess how the inclusion of supplementary thermal data could enhance the accuracy of mineral volume fraction predictions. Fig. 10 displays the comparative predictions of mineral volume fractions for the first well, both with and without the incorporation of thermal data. The predictions exhibit enhancement with the incorporation of supplementary thermal profiling data. Specifically, the RMSE decreases to 0.039 when thermal data is included, as opposed to 0.046 without its inclusion. With the necessary input data and model selection criteria in place for achieving accurate predictions of mass and volume mineral composition, albeit with certain variations at specific intervals, the focus shifts to assessing the model's efficacy beyond evaluation solely based on RMSE and MAE. To explore the model's performance more comprehensively and link the rock's thermal component with mineralogy in unconventional reservoirs, predictions were extended to a sub-task involving the calculation of thermal conductivity. The densities and thermal conductivities (TC) of individual minerals were optimized only within the hold-out calibration intervals by minimizing the discrepancy between the experimental TC (measured via Optical Scanning) and the TC calculated from Litho Scanner-based mineral compositions (Fig. 11). The optimization was constrained within reference ranges of literature values and did not use any intervals from the test well, ensuring the inde-

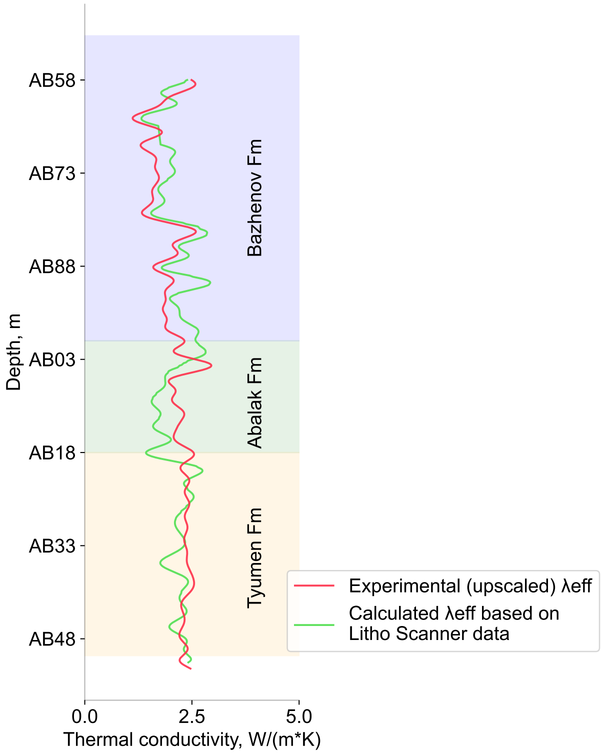


Fig. 11. Parallel thermal conductivity from core measurements vs calculated thermal conductivity from original volume fractions of minerals based on calibrated mineral densities on a train well.

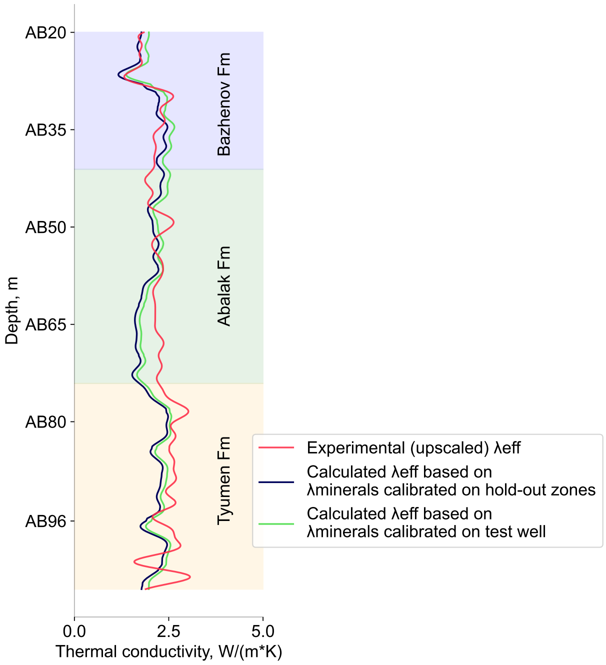


Fig. 12. Parallel thermal conductivity from core measurements vs calculated thermal conductivity from predicted volume fractions based on calibrated TC of minerals on hold-out zones/test well.

pendence of the physical parameter fitting from model validation. The obtained mineral parameters (Table 7) were then applied unchanged to predict TC on the held-out test well, allowing for a fully independent physics-based comparison. For a blind-well three configurations were considered: Calibrated TC of minerals on a hold-out zones, test-calibrated TC of minerals fitted on the test well to minimize RMSE (reported only as an upper bound) and laboratory values of effective TC, see Fig. 12.

The results confirm that the calibration performed on hold-out intervals is transferable: The agreement between predicted and measured TC remains consistent across the blind test well.

To evaluate the robustness of the thermal-conductivity calculation with respect to uncertainties in mineral properties, we performed a sensitivity analysis varying thermal conductivities of key minerals within their reference ranges. Fig. 13 presents a tornado plot illustrating the influence of ±10 % variations in each parameter on the resulting effective thermal conductivity λ_{eff}. The analysis reveals that the model is most sensitive to the thermal conductivities of clay and quartz. However, even under $\pm 10\%$ variations within geologically plausible limits, the RMSE between the calculated and experimental λ_{eff} did not exceed 0.1 W/(m*K), confirming the robustness of the proposed calibration and the independence of the physics-based verification step.

4. Discussions and conclusions

A novel approach was developed for determining the mineralogical composition of the rock structure in the West Siberian oil field and the Bazhenov Formation, which is associated with unconventional reservoirs, using conventional well logs and thermal profiling data. The relationship between mineralogy and the thermal component of the rock was validated through subsequent calculation of thermal conductivity based on a proven theoretical model. The research involved applying a tailored methodology for data preparation and integration, selecting appropriate models, and optimizing hyperparameters, while comparing the effectiveness of each algorithm.

The evaluation of the model utilized Litho Scanner measurements as target values, representing the mineral content (w/w) comprising clay, calcite, dolomite, pyrite, coal, quartz-feldspar-mica, and siderite. Various prediction strategies were developed using different datasets and models (KNeighbors, Random Forest, LightGBM, XGBoost, CatBoost, GradientBoostingRegressor) organized into Multioutput Regressor and Regressor Chain frameworks. Despite comparable numerical performance of other tree-based baselines (CatBoost, LightGBM, XGBoost), the GBR + RegressorChain model provided more stable predictions across depth-blocked folds and superior interpretability, making it the most suitable choice for the compositional regression task given the limited dataset.

This is particularly important for unconventional reservoirs, where certain minerals may influence the formation and deposition of others. To ensure that the model appropriately weights significant variables, such as thermal properties and conventional well logs suitable for mineralogy determination, the assessment of feature importance was conducted using the SHAP technique. Based on these findings, the input well logs were selected with consideration of standard well measurements typically conducted in the Bazhenov Formation. Consequently, the following logging curves were utilized as

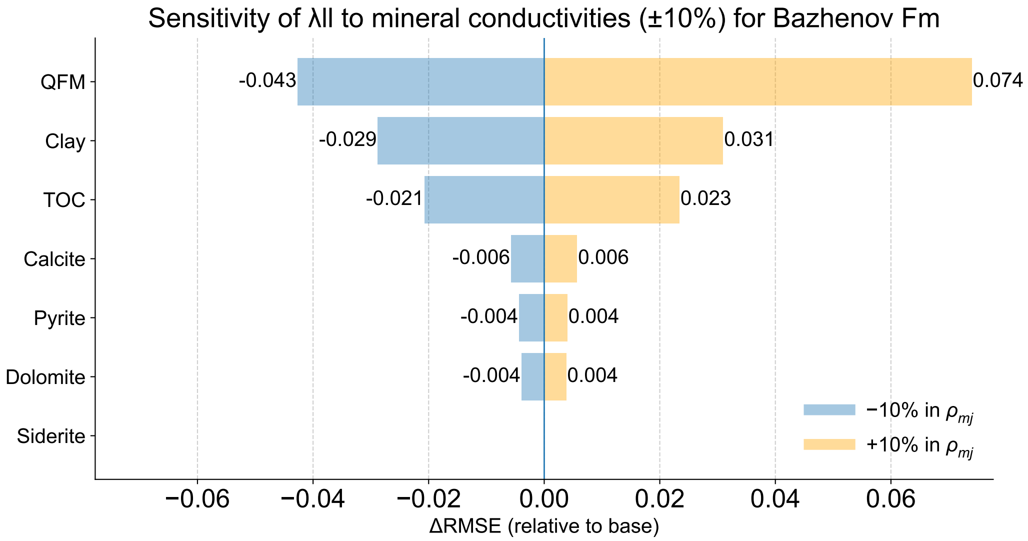


Fig. 13. The tornado plot illustrates the influence of ±10 % variations in TC of minerals on the accuracy of total TC prediction.

input: GR, RHOZ, RXOZ, HCAL, TNPH, PEFZ, DTCO, DTSM, and BMK. Additionally, the mineral fractions were found to have a significant impact on thermal properties. This was demonstrated by assessing Shapley values, where the perpendicular component of thermal conductivity was among the top 9 and top 12 important features for quartz and pyrite, respectively. To avoid overfitting, GridSearch was utilized with 5-fold cross-validation; in addition, GradientBoostingRegressor incorporates built-in regularization mechanisms by controlling the learning rate, which was set to 0.01 after grid search optimization.

While the model is data-driven, the selection of input features along with thermal conductivities of minerals was guided by geological knowledge, particularly the importance of these parameters in predicting mineral distributions in the Bazhenov Formation.

Several strategies for combining input data were developed. In cases of data scarcity, more robust approaches, such as data augmentation, are required. However, the GradientBoostingRegressor wrapped in a RegressorChain proved to be the most stable in terms of RMSE and MAE. Dry-weight fractions were accurately predicted, with the best RMSE reaching 0.026 for strategies involving two training wells and one blind well. This result demonstrates that using conventional well logs as input features enables the model to capture trends in mineral weight fractions effectively in highly heterogeneous reservoirs. The ablation results confirm that the inclusion of thermal-related logs does not contribute significantly to predictive accuracy of dry-weight fractions.

The conversion of mineral weight fractions to volume fractions was performed using the TOC parameter, which, along with certain minerals, is strongly correlated with the rock's thermal properties. Incorporating thermal profiling data alongside conventional logging data as input to the model resulted in improved predictions of mineral volume fractions, reducing the RMSE from 0.046 to 0.039.

Furthermore, the outcomes derived from the model predictions (with additional thermal data), i.e., volumetric mineral fractions, were used in the theoretical model for TC calculation. The comparison of calculated and measured TC values revealed that the ML models exhibited substantial accuracy, indicating their suitability for leveraging mineral component predictions in reservoir characterization efforts.

Moreover, separating the calibration of mineral densities and thermal conductivities onto dedicated hold-out intervals and validating on a blind test well ensured that the physics-based verification remained independent of the ML training process. The additional sensitivity analysis demonstrated that even substantial parameter perturbations within reference ranges result in minimal changes (δRMSE < 0.1 W/m*K), confirming the overall robustness and generalizability of the workflow.

When rocks contain numerous minerals with variable sizes, textures, and structures, well logging data combined with thermal profiling are sufficient for accurate mineralogy prediction and utilization of results for subsequent subtasks. However, it is suggested to use additional data such as XRD/XRF, which have higher resolution and can provide valuable insights into mineral component modeling, particularly when data are scarce. Subsequent efforts will focus on refining the model's performance, particularly for clay and quartz, with the objective of identifying inherent patterns within heterogeneous and anisotropic rock formations.

The developed approach requires validation on other unconventional reservoirs with similar rock properties and comparable input variables. Caution is advised when applying this method to estimate mineralogy at new sites, as mineral types and associations may vary due to differences in climatic conditions, local geology, and primary sediment sources. For cases of data scarcity, it is suggested to use data augmentation, which was not implemented in this study.

The application of the developed model has practical significance for geological modeling, facies analysis, reservoir rock characterization, and optimization of enhanced oil recovery methods. Future applications could include real-time decision-making during drilling, identifying sweet spots in unconventional formations by analyzing mineralogy variations in rocks based on logging curves typically available during well measurements in West Siberian oil fields. This is beneficial for both reservoir engineers and petrophysicists and could reduce economic costs associated with additional well-logging operations. For petroleum engineers, the focus should be on how the model enhances geological understanding and supports field operations. Moreover, it highlights the innovative application of the GradientBoostingRegressor with RegressorChain and how it can be adapted to other fields dealing with heterogeneous and sparse data.

Acknowledgements

This work was supported by the Ministry of Science and Higher Education of the Russian Federation under agreement (No. 075-10-2022-011) within the framework of the development program for a world-class Research Center.

Additional information: Author's email

D.Koroteev@skoltech.ru (D. Koroteev); Y.Popov@skoltech.ru (Y. Popov).

Conflicts of interest

The authors declare no competing interest.

References

Alekseev, A. D., Gavrilov, A. E. Methodical bases for the construction of integrated petrophysical models of unconventional and complex reservoirs based on the special core analysis results. PROneft Journal, 2019, 3: 2534. (in Russian).

Amosova, A., Panteeva, S., Tatarinov, V., et al. X-ray fluorescence determination of major rock-forming elements in small samples 50 and 110 mg. Analitika i Kontrol', 2015, 19(2): 130-138. (in Russian).

Barham, A., Zainal Abidin, N. Machine learning approach to predict the illite weight percent of unconventional reservoirs from well-log data: An example from Montney Formation, NE British Columbia, Canada. Applied Sciences, 2023, 14: 318.

Barshad, I. Thermal analysis techniques for mineral identification and mineralogical composition. Methods of Soil Analysis: Part 1. Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling, 1965, 9: 699-742.

Beck, A. E., Darbha, D. M., Schloessin, H. H. Lattice conductivities of single-crystal and polycrystalline materials at mantle pressures and temperatures. Physics of the Earth and Planetary Interiors, 1978, 17(1): 35-53.

Breiman, L. Random forests. Machine Learning, 2001, 45: 5-32.

Brigham, E. O. The Fast Fourier Transform and Its Applications. Prentice-Hall, Englewood Cliffs, USA, 1988.

Chai, T., Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding rmse in the literature. Geoscientific Model Development, 2014, 7: 1247-1250.

Chekhonin, E., Popov, Y., Romushkevich, R., et al. Integration of thermal core profiling and scratch testing for the study of unconventional reservoirs. Geosciences, 2021, 11(6): 260.

Clauser, C., Huenges, E. Thermal conductivity of rocks and minerals, in Rock Physics and Phase Relations: A Handbook of Physical Constants, edited by T. J. Ahrens, American Geophysical Union, Washington, DC, USA, pp. 105-126, 1995.

Conn, A. R., Gould, N. I. M., Toint, P. L. Trust Region Methods. Society for Industrial and Applied Mathematics, Philadelphia, USA, 2000.

Craddock, P., Srivastava, P., Datir, H., et al. Enhanced mineral quantification and uncertainty analysis from downhole spectroscopy logs using variational autoencoders. Paper SPWLA 2021-v62n6a2 Presented at Proceedings of the SPWLA Annual Symposium, Virtual, 17-20 May, 2021.

Cui, Q., Yang, H., Li, X., et al. Identification of lithofacies and prediction of mineral composition in shales: A case study of the shahejie formation in the bozhong sag. Unconventional Resources, 2022, 2: 72-84.

Dang, S. T., Sondergeld, C. H., Rai, C. S. A new approach to measuring organic density. Petrophysics, 2016, 57(2): 112-120.

Gavrilov, A. E., Zhukovskaya, E. A., Tugarova, M. A., et al. Objective bazhenov rocks classification (the case of the west siberia central part fields). Neftyanoe Khozyaystvo – Oil Industry, 2015, 12: 38-40. (in Russian).

Hall, P., Park, B. U., Samworth, R. J. Choice of neighbor order in nearest-neighbor classification. The Annals of Statistics, 2008, 36(5): 2135-2152.

Hantschel, T., Kauerauf, A. I. Fundamentals of Basin and Petroleum Systems Modeling. Springer, Berlin, Germany, 2009.

Hastie, T., Tibshirani, R., Friedman, J. H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, USA, 2009.

Horai, K.-i. Thermal conductivity of rock-forming minerals. Journal of Geophysical Research, 1971, 76(5): 1278-1308.

Hu, K., Liu, X., Chen, Z., et al. Mineralogical characterization from geophysical well logs using a machine learning approach: Case study for the Horn River Basin, Canada. Earth and Space Science, 2023, 10: e2023EA003084.

Khan, S. Q., Kirmani, F. Applicability of deep neural networks for lithofacies classification from conventional well logs: An integrated approach. Petroleum Research, 2024, 9: 393-408.

Kim, D., Choi, J., Kim, D., et al. Predicting mineralogy by integrating core and well log data using a deep neural network. Journal of Petroleum Science and Engineering, 2020, 195: 107838.

Kodikara, G. R. L., McHenry, L. J., Stanistreet, I. G., et al. Wide and deep learning for predicting relative mineral compositions of sediment cores solely based on xrf scans: A case study from pleistocene paleolake olduvai, tanzania. Artificial Intelligence in Geosciences, 2024, 5: 100088.

Kozlov, E. N., Fomina, E. N. The use of factor analysis for express diagnostics of the mineral composition of geologically complex objects based on X-ray diffraction data. Paper Presented at Proceedings of the Geological Institute of the Russian Academy of Sciences, Moscow, Russia, 20-23 November, 2018. (in Russian).

Kumar, T., Seelam, N. K., Rao, G. S. Lithology prediction from well log data using machine learning techniques: A case study from talcher coalfield, eastern india. Journal of Applied Geophysics, 2022, 199: 104605.

Martin, T., Meyer, R., Jobe, Z. Centimeter-scale lithology and facies prediction in cored wells using machine learning. Frontiers in Earth Science, 2021, 9: 659611.

Meshalkin, Y., Shakirov, A., Orlov, D., et al. Well-logging based lithology prediction using machine learning. Paper Prsented at Proceedings of the EAGE Conference on Data Science in Oil and Gas, Vienna, Austria, 19-20 October, 2020.

Nawal, M., Shekar, B., Jaiswal, P. Integration of sparse and continuous data sets using machine learning for core mineralogy interpretation. The Leading Edge, 2023, 42: 421-432.

Park, S., Son, B.-K., Choi, J., et al. Application of machine learning to quantification of mineral composition on gas hydrate-bearing sediments, ulleung basin, korea. Journal of Petroleum Science and Engineering, 2021, 209: 109840.

Popov, Y., Beardsmore, G., Clauser, C., et al. ISRM suggested methods for determining thermal properties of rocks from laboratory tests at atmospheric pressure. Rock Mechanics and Rock Engineering, 2016, 49(10): 4179-4207.

Popov, Y., Berezin, V., Soloviov, V., et al. Thermal conductivity of minerals. Izvestia, Physics of Solid Earth, 1987, 23(3): 245-253.

Postnikova, O. V., Postnikov, A. V., Zueva, O. A., et al. Types of void space in the bazhenov reservoir rocks. Geosciences, 2021, 11(7): 269.

Rodriguez-Galiano, V., Sánchez-Castillo, M., Chica-Olmo, M., et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 2015, 71: 804-818.

Sass, J. H. The thermal conductivity of fifteen feldspar specimens. Journal of Geophysical Research, 1965, 70(16): 4064-4065.

Serra, O. Fundamentals of Well-Log Interpretation. Elsevier, Amsterdam, Netherlands, 1983.

Temnikova, E. Y., Fedoseev, A. A., Kazanenkov, V. A., et al. Lithological characteristic of bazhenov formation sections in central and southeastern regions of western siberia according to logging data set. Russian Geology and Geophysics, 2022, 63(9): 1050-1060.

Tucker, M. Techniques in Sedimentology. Blackwell Science, Oxford, UK, 1988.

Tóth, C., Harangi, S., Károlyi, A., et al. Method development for the elemental analysis of organic rich soil samples by microwave plasma atomic emission spectrometry. Studia Universitatis Babes-Bolyai Chemia, 2017, 62: 483-494.

Wang, W., Carreira-Perpinan, M. A. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. ArXiv Preprint ArXiv: 1309.1541, 2013.

Yang, Z., Ghanizadeh, A., Younis, A., et al. Prediction of mineralogical composition in heterogeneous unconventional reservoirs: Comparisons between data-driven and chemistry-based models. Paper SPE 218116 D011S004R001 Prsented at Proceedings of the SPE Canadian Energy Technology Conference, Calgary, Canada, 13-14 March, 2024.

Zhou, B., Luo, R., Li, N., et al. Predicting mineralogy of the salt-gypsum layer by drilling cuttings and conventional well logging data using a multilayer perceptron neural network (MLPNN): A case study in kuqa depression, tarim basin. Paper ARMA 2021-1418 Prsented at 55th U.S. Rock Mechanics/Geomechanics Symposium, Virtual, 18-25 June, 2021.