ISSN: 2688-8645

Well-Grouped Validation and Explainable XGBoost Lithology Classification under Missing Well-Log Conditions

Haowei Li [1, 2, 3]

[1] School of Petroleum Engineering, Xi'an Shiyou University, Xi'an, Shaanxi, 710065, China

[2] Engineering Research Center of Development and Management for Low to Ultra-Low Permeability Oil & Gas Reservoirs in West China, Ministry of Education, Xi'an, Shaanxi, 710065, China

[3] Key Laboratory of Exploration and Development of Complex and Difficult-to-Produce Oil & Gas Reservoirs, Ministry of Education, Xi'an, Shaanxi, 710065, China

Abstract

To address challenges in actual logging interpretation—such as imbalanced lithology types, incomplete logging curves, and the difficulty of reliably evaluating cross-well generalization—this study performs intelligent lithology identification using the publicly available FORCE 2020 logging dataset. After preprocessing, the dataset comprises 98 wells, 1,170,511 depth samples, and 12 lithology types. Five conventional logging curves—natural gamma ray, bulk density, neutron porosity, acoustic transit time (DTC), and deep laterolog resistivity (RDEP)—are selected as input features. A 5-fold GroupKFold cross-validation ensures complete separation of training and test wells. XGBoost, LightGBM, and random forest classifiers are compared, together with various missing-value handling and class-weighting strategies. On this basis, model robustness to incomplete logging data is evaluated under three missing-data scenarios: random point missing, continuous interval missing, and entire-curve omission. For the entire-curve case, a compromise model structure is adopted. Out-of-fold predictions are used for per-class error diagnosis, and TreeSHAP is applied to interpret the model's decision rationale. The results show that XGBoost, using native missing-value handling and no class weighting, achieves the most balanced cross-well performance, with an accuracy of 0.687, a Macro-F1 of 0.328, and an MCC of 0.430. Introducing 30% random missing points reduced Macro-F1 by 36.98% and MCC by 39.63%, indicating that class-balance metrics are more sensitive than overall accuracy. Considerable confusion exists between the sandstone–shale transitional lithology and pure shale; 67.02% of true transitional samples are misclassified as shale. SHAP analysis of class-wise contributions reveals that RDEP, RHOB, and DTC are the most important logging curves overall, and feature dependence varies significantly among different lithologies. These findings can serve as a reference for batch preliminary lithology screening, logging quality control, and manual review when logging data are incomplete.

Keywords

Lithology Identification; Well-level Cross-validation; Logging Curves; XGBoost; Class Imbalance; Out-of-fold Prediction; SHAP.

1. Introduction

Lithology identification is a fundamental task in well-log interpretation and reservoir evaluation, directly affecting reservoir delineation, petrophysical interpretation, sedimentary facies analysis, and geological modeling [1,2]. Conventional interpretation relies on the combined responses of gamma ray, resistivity, density, neutron, and acoustic logs together with

ISSN: 2688-8645

geological knowledge and interpreter experience [3,4]. However, logging responses vary across wells, burial conditions, and diagenetic settings, while different lithologies may exhibit overlapping responses because of similar mineral composition, pore structure, and fluid properties [5]. These factors make manual interpretation time-consuming and susceptible to subjective judgment and data quality [6,7].

With the accumulation of digital well-log data, machine-learning methods have been increasingly applied to lithology and lithofacies classification, including support vector machines, random forests, gradient boosting models [8-11], convolutional neural networks [12-14], and recurrent neural networks [15-17]. These methods can capture nonlinear relationships among multiple logging curves and improve batch interpretation efficiency [18-20]. More recent studies have extended intelligent lithology interpretation toward uncertainty analysis, blind-well prediction, and imbalanced or small-sample learning [21-27]. Accordingly, the key issue is gradually shifting from fitting known wells to achieving reliable generalization, robustness, and interpretability in unseen wells.

Three challenges remain particularly important. First, random depth-sample splitting may place neighboring samples from the same well in both training and test sets, leading to overly optimistic results and failing to reflect true cross-well generalization [11,22,28]. Differences in logging baselines, lithology proportions, and local response patterns among wells further increase the difficulty of unseen-well prediction [29-32]. Second, practical logging data frequently contain isolated missing values, continuous bad intervals, or completely absent curves due to tool failure, borehole conditions, incomplete logging suites, or quality-control rejection [33-35]. Although previous studies have focused on missing-data reconstruction and imputation, the effects of different missing patterns on lithology classification, particularly on minority classes, remain insufficiently quantified. Third, severe class imbalance can bias global feature-importance analysis toward dominant lithologies [36-38]. Although SHAP provides a useful framework for model interpretation, its global importance results are still influenced by the composition of the explanation samples [39-41].

To address these issues, this study develops a lithology-classification workflow for unseen wells and incomplete logging conditions. Specifically, (1) well-grouped five-fold cross-validation is used to ensure complete separation between training and test wells and to compare tree-based models and preprocessing strategies; (2) random point missing, continuous interval missing, and entire-curve omission are introduced only into the test folds to evaluate robustness under different data-loss scenarios; and (3) OOF error diagnosis is combined with class-balanced TreeSHAP to identify major failure modes and lithology-specific feature dependence. The proposed workflow is intended to support cross-well lithology pre-screening, logging quality control, and manual review under incomplete logging conditions.

2. Data and Methods

2.1. Logging Data and Lithology Distribution

This study uses the FORCE 2020 well-log and lithofacies dataset from the Norwegian Sea. After field checking, invalid-record processing, and feature selection, the dataset contains 98 wells, 1,170,511 depth samples, and 12 lithology classes: sandstone, shale, sandstone/shale, limestone, chalk, dolomite, marl, anhydrite, halite, coal, basement rock, and tuff. All depth samples are retained for the experiments[18]. The overall workflow is shown in Fig. 1.

Five conventional logging curves are selected as model inputs: natural gamma ray (GR), bulk density (RHOB), neutron porosity (NPHI), acoustic transit time (DTC), and deep laterolog resistivity (RDEP). These curves provide complementary information on shale content, matrix density, apparent porosity, acoustic properties, pore fluids, and formation conductivity. Their physical meanings and missing-data rates are summarized in Table 1.

ISSN: 2688-8645

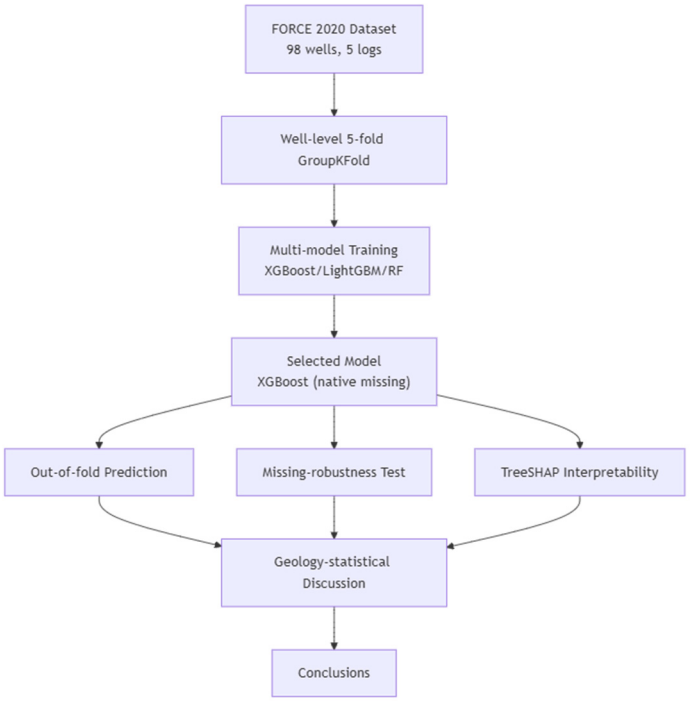

Fig 1. Research technical roadmap.

Table 1. Input logging curves and missing data status

Logging curve	Physical meaning	Missing rate / %	Included in main model
GR	Natural radioactivity and shale content response	0.00	Yes
RHOB	Formation bulk density	13.78	Yes
NPHI	Hydrogen index and apparent neutron porosity	34.61	Yes
DTC	Compressional acoustic transit time	6.91	Yes
RDEP	Deep formation resistivity	0.94	Yes
PEF	Photoelectric absorption cross-section index	42.62	No

The photoelectric factor (PEF) is sensitive to mineral composition and may provide additional information for distinguishing mineralogically similar lithologies. However, its overall missing rate reaches 42.62%, and it is completely absent from 29 wells. Considering that the present study focuses on cross-well lithology classification under incomplete conventional logging conditions, PEF is not included in the main model. Its potential contribution can be further evaluated when more complete mineral-sensitive logging data are available.

The dataset exhibits pronounced class imbalance. Shale accounts for 720,803 samples, corresponding to 61.58% of the total dataset, while sandstone and sandstone/shale account for 14.43% and 12.85%, respectively. Several lithologies, including dolomite, anhydrite, halite, coal, and basement rock, contain substantially fewer samples. The sample size and well coverage of each lithology are listed in Table 2.

For lithologies occurring in only one or a few wells, well-wise data splitting may result in a training fold that contains very limited examples or no examples of a particular class. Consequently, poor recognition of these classes cannot be attributed solely to the classification algorithm.

ISSN: 2688-8645

Table 2. Sample size and well coverage of each lithology

Lithology	Number of samples	Percentage/%	Number of wells
Sandstone	168937	14.43	98
Shale	720803	61.58	98
Sandstone/shale	150455	12.85	97
Limestone	56320	4.81	94
Chalk	10513	0.90	11
Dolomite	1688	0.14	37
Marl	33329	2.85	73
Anhydrite	1085	0.09	6
Halite	8213	0.70	3
Coal	3820	0.33	51
Basement rock	103	0.01	1
Tuff	15245	1.30	98

2.2. Well-grouped Validation and Model Comparison

To prevent information leakage between training and test samples from the same well, a five-fold GroupKFold cross-validation scheme is adopted. The 98 wells are divided into five mutually exclusive subsets. In each fold, one subset is used as the test set and the remaining subsets are used for model training. Each well therefore belongs entirely to either the training or test portion of a given fold, and depth samples from the same well never appear in both simultaneously.

This design ensures that all test samples originate from wells unseen during model fitting and more closely represents the practical application of a trained model to newly logged wells. It also reduces the risk that the model exploits well-specific curve baselines, calibration characteristics, or similarity between adjacent depth samples.

Model performance is evaluated using Accuracy, Balanced Accuracy, Macro-F1, Weighted-F1, and the multiclass Matthews Correlation Coefficient (MCC). Accuracy and Weighted-F1 reflect overall sample-level classification performance, whereas Balanced Accuracy and Macro-F1 assign greater importance to performance across individual classes. MCC incorporates the overall confusion structure and provides a complementary measure under class-imbalanced conditions. Because shale accounts for more than 60% of the dataset, overall Accuracy alone is insufficient to evaluate minority-lithology performance. The standard mathematical definitions of these commonly used metrics are omitted for brevity. The original chapter contained full definitions for all five metrics.

Three tree-based classifiers are compared: eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF). XGBoost and LightGBM can directly handle missing values by learning default branch directions during tree construction, whereas RF is evaluated using median imputation with missing-indicator variables. In the imputation scheme, the median of each feature is calculated using only the current training fold and is then used to impute both the training fold and the corresponding test fold. A binary missing indicator is simultaneously added to preserve information related to the original missing status.

For XGBoost, both native missing-value handling and median imputation with missing indicators are evaluated. To further examine the effect of class imbalance, XGBoost with native missing-value handling is also tested using inverse-frequency class weighting. The weight assigned to class k is defined as:

$$w_k = \frac{N_{\mathrm{tr}}}{K_{\mathrm{tr}} n_{k,\mathrm{tr}}} \tag{1}$$

ISSN: 2688-8645

where N_{tr} denotes the total number of samples in the current training fold, K_{tr} is the number of lithology classes present in that fold, and $n_{k,tr}$ is the number of training samples belonging to class k.

The comparison therefore includes XGBoost with native missing-value handling, XGBoost with median imputation and missing indicators, LightGBM with native missing-value handling, RF with median imputation and missing indicators, and XGBoost with inverse-frequency class weighting. The model showing the best overall balance among the evaluation metrics is subsequently used for missing-data robustness tests, OOF error diagnosis, and SHAP interpretation.

2.3. Missing Data Robustness Experimental Design

To evaluate model robustness to incomplete logging data encountered in new wells, synthetic missing values are introduced only into the test portion of each cross-validation fold, while the corresponding training fold retains its original natural missingness. This design prevents the model from adapting to the artificially generated missing patterns during training, so that changes in test performance primarily reflect robustness to additional information loss in unseen wells.

Three missing-data scenarios are considered.

(1) Random point missing:A specified proportion of originally valid values among the five input curves is randomly selected and set to missing in each test fold. Missing ratios of 10% and 30% are considered. This scenario represents scattered measurement anomalies, data transmission loss, or isolated quality-control rejection.

(2) Continuous interval missing: Continuous missing intervals are generated along the depth sequence of the test wells, with a nominal missing proportion of 30%. This scenario represents logging-tool malfunction, deterioration of borehole conditions, or unusable measurements over continuous depth intervals.

(3) Entire curve omission: In each fold, 30% of the test wells are selected and either the complete NPHI or RHOB curve is removed. Independent experiments are conducted for NPHI and RHOB using the same test-well selection scheme. This scenario represents incomplete historical logging suites, differences in tool combinations, or complete rejection of a logging curve during quality control.

The three scenarios represent different patterns of information loss and should not be interpreted as strictly equivalent missing-data budgets. In particular, random point missing, continuous interval missing, and entire-curve omission affect different numbers of depth samples and different combinations of available curves. Their results are therefore compared primarily in terms of degradation trends rather than as strictly equivalent perturbations.

2.4. OOF Diagnosis and TreeSHAP Methodology

During five-fold cross-validation, the true lithology labels, predicted labels, well identifiers, depths, and maximum predicted class probabilities of all test samples are retained. Because each sample is predicted only when its corresponding well belongs to the test fold, predictions from the five folds can be concatenated to obtain complete out-of-fold (OOF) predictions for the entire dataset.

The OOF predictions are used to calculate class-wise precision, recall, and F1-score, as well as the row-normalized confusion matrix, major directional misclassification paths, and per-well performance. Prediction confidence is also analyzed. A misclassified sample with a maximum predicted class probability of at least 0.90 is defined as a high-confidence error. This analysis is used to distinguish uncertain predictions from systematic high-confidence misclassification.

TreeSHAP is applied to interpret the selected XGBoost model. SHAP values quantify the contribution of individual logging features to the model output and allow differences in feature

ISSN: 2688-8645

dependence among lithology classes to be examined. Because the dataset is strongly imbalanced, two explanation-sampling strategies are adopted.

Under original-distribution sampling, explanation samples are selected according to their occurrence in the complete dataset. The resulting global importance therefore represents model feature dependence under the actual lithology composition but is strongly influenced by the dominant shale class.

Under class-balanced sampling, equal numbers of samples are selected from each of the 12 lithology classes. Each lithology therefore contributes equally to the global SHAP importance calculation, reducing the dominance of high-frequency classes and improving the visibility of feature dependence for minority lithologies.

In addition, stratified bootstrap resampling is performed on the class-balanced explanation samples to assess the stability of the global feature-importance ranking with respect to variation in the explanation sample. SHAP values are further grouped according to the true lithology labels to investigate class-specific dependence on GR, RHOB, NPHI, DTC, and RDEP.

3. Results and Discussion

3.1. Comparison of Models and Processing Strategies

The five-fold GroupKFold results for different models and processing strategies are summarized in Table 3 and Fig. 2. XGBoost with native missing-value handling and no class weighting achieves the most balanced overall performance, with an Accuracy of 0.687, Balanced Accuracy of 0.309, Macro-F1 of 0.328, Weighted-F1 of 0.654, and MCC of 0.430. XGBoost with median imputation and missing indicators gives a slightly higher Accuracy of 0.689, but its Balanced Accuracy and Macro-F1 decrease to 0.297 and 0.312, respectively. The differences between the two XGBoost missing-value schemes are therefore limited, and native missing handling is preferred mainly because of its simpler workflow and relatively better class-balanced performance.

Table 3. Cross-well cross-validation performance of different models and processing strategies

Model & strategy	Accuracy	Balanced Accuracy	Macro-F1	Weighted-F1	MCC
XGBoost, native missing, no weighting	0.687±0.027	0.309±0.055	0.328±0.048	0.654±0.035	0.430±0.043
LightGBM, native missing, no weighting	0.667±0.030	0.307±0.046	0.300±0.040	0.644±0.032	0.400±0.043
Random Forest, median + missing indicator	0.675±0.036	0.293±0.052	0.306±0.040	0.645±0.035	0.412±0.056
XGBoost, median + missing indicator	0.689±0.028	0.297±0.045	0.312±0.034	0.654±0.038	0.430±0.046
XGBoost, native missing, class weighting	0.490±0.056	0.385±0.038	0.269±0.019	0.549±0.051	0.322±0.044

LightGBM and RF show slightly lower Macro-F1 and MCC than XGBoost under the same feature set and well-level validation framework. More importantly, inverse-frequency class weighting increases Balanced Accuracy from 0.309 to 0.385 but reduces Accuracy to 0.490 and also lowers Macro-F1, Weighted-F1, and MCC. This indicates that simple inverse-frequency weighting improves minority-class recall at the cost of overall prediction reliability. Under severe class imbalance, excessive weights assigned to very rare classes can shift a substantial number of majority-class samples toward minority labels.

ISSN: 2688-8645

Considering both overall stability and minority-class performance, XGBoost with native missing-value handling and no class weighting is selected as the main model for subsequent OOF diagnosis, missing-data robustness analysis, and SHAP interpretation.

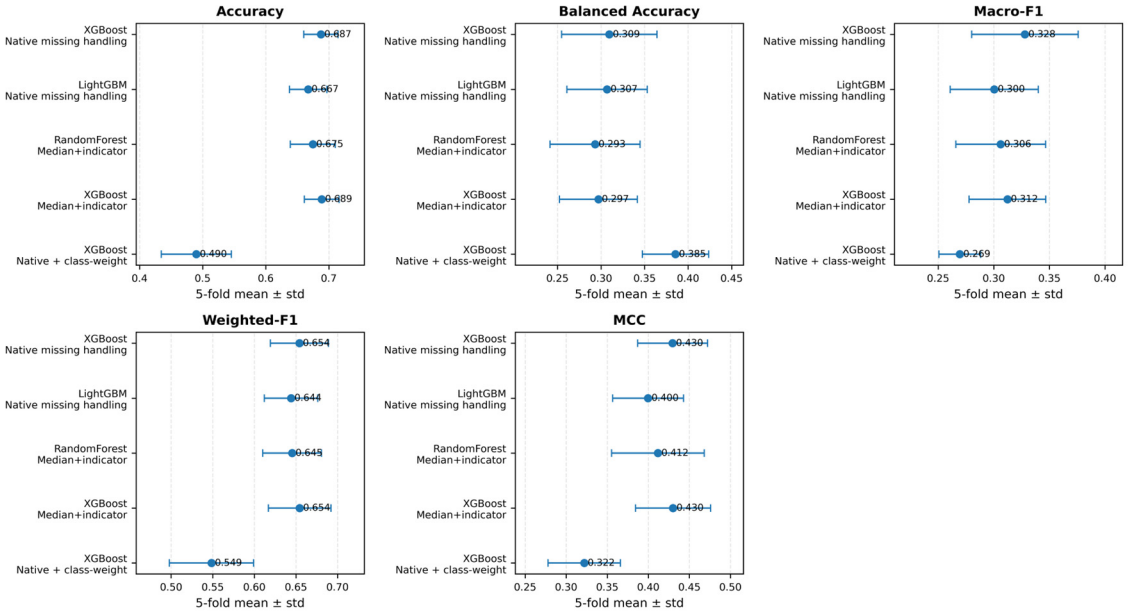

Fig 2. Comparison of five-fold performance among different models and processing strategies.

3.2. Per-class Performance and Misclassification Patterns

The OOF class-wise performance of the selected XGBoost model is shown in Table 4, and the row-normalized confusion matrix is presented in Fig. 3.

Table 4. Per-lithology OOF performance of the XGBoost main model

Lithology	Precision	Recall	F1	Samples
Sandstone	0.582	0.654	0.616	168937
Shale	0.776	0.885	0.827	720803
Sandstone/shale	0.238	0.124	0.163	150455
Limestone	0.565	0.485	0.522	56320
Chalk	0.215	0.103	0.140	10513
Dolomite	0.119	0.017	0.030	1688
Marl	0.260	0.124	0.168	33329
Anhydrite	0.771	0.701	0.735	1085
Halite	0.945	0.088	0.161	8213
Coal	0.539	0.409	0.465	3820
Basement rock	0.000	0.000	0.000	103
Tuff	0.294	0.104	0.153	15245

Shale is the best-identified major lithology, with an F1-score of 0.827, followed by sandstone (0.616) and limestone (0.522). Anhydrite reaches an F1-score of 0.735 despite having only 1,085 samples, suggesting that class separability depends not only on sample size but also on whether the lithology exhibits a distinctive multi-curve response. In contrast, several transitional or mineralogically similar classes remain difficult to identify.

The dominant error path is sandstone/shale → shale. The sandstone/shale class has a recall of only 0.124, and 67.02% of its true samples, corresponding to 100,837 depth samples, are predicted as shale. This single pathway accounts for 27.56% of all OOF errors. The main reason

ISSN: 2688-8645

is that sandstone/shale represents a transitional lithology whose logging responses vary continuously with sandy and argillaceous composition, whereas the target labels are discrete. Combined with the dominant proportion of shale in the training data, transitional samples are therefore easily absorbed into the shale decision region.

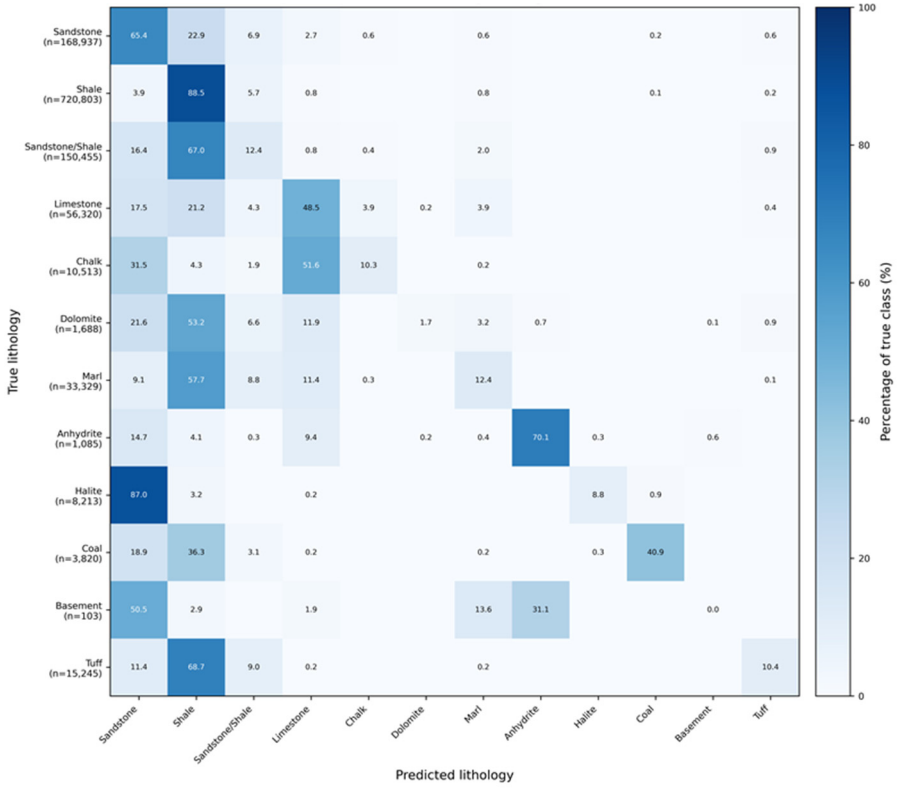

Fig 3. Row-normalized OOF confusion matrix of the XGBoost main model.

Other poor-performing classes exhibit different limitations. Dolomite has an F1-score of only 0.030, reflecting limited samples and substantial response overlap with limestone and other dense carbonates when mineral-sensitive measurements such as PEF are unavailable. Basement rock has only 103 samples from a single well; when that well is assigned to the test fold, no basement-rock examples remain in training, making meaningful recognition impossible. Halite, in contrast, shows very high precision (0.945) but low recall (0.088), indicating that only its most typical logging signatures are reliably identified.

High-confidence errors further reveal a systematic bias toward shale. Among the 33,312 incorrect predictions with confidence ≥ 0.90, 78.83% are predicted as shale, and sandstone/shale → shale alone accounts for 42.88% of these errors. This suggests that some errors are not merely uncertain boundary cases but confident assignments toward the dominant class.

Performance also varies markedly among wells. The median well-level Accuracy is 0.721, ranging from 0.199 to 0.937, and 10 wells have an Accuracy below 0.50. Therefore, global average metrics cannot fully represent model behavior in every well, particularly where lithology proportions, curve baselines, or missing-data patterns differ substantially from the training population.

3.3. Robustness to Missing Logging Data

The results under different synthetic missing-data scenarios are summarized in Table 5 and Fig. 4.

ISSN: 2688-8645

Table 5. Model performance under different missing scenarios in test folds

Testing scenario	Accuracy	Balanced Accuracy	Macro-F1	Weighted-F1	MCC
Natural missing	0.687±0.027	0.309±0.055	0.328±0.048	0.654±0.035	0.430±0.043
10% random point missing	0.665±0.024	0.263±0.038	0.288±0.038	0.626±0.034	0.368±0.037
30% random point missing	0.629±0.027	0.183±0.021	0.207±0.027	0.577±0.039	0.259±0.034
30% continuous interval missing	0.637±0.029	0.201±0.052	0.221±0.053	0.587±0.042	0.287±0.054
Entire NPHI missing in 30% test wells	0.677±0.029	0.280±0.053	0.302±0.056	0.645±0.037	0.406±0.048
Entire RHOB missing in 30% test wells	0.675±0.032	0.273±0.052	0.297±0.055	0.637±0.038	0.396±0.059

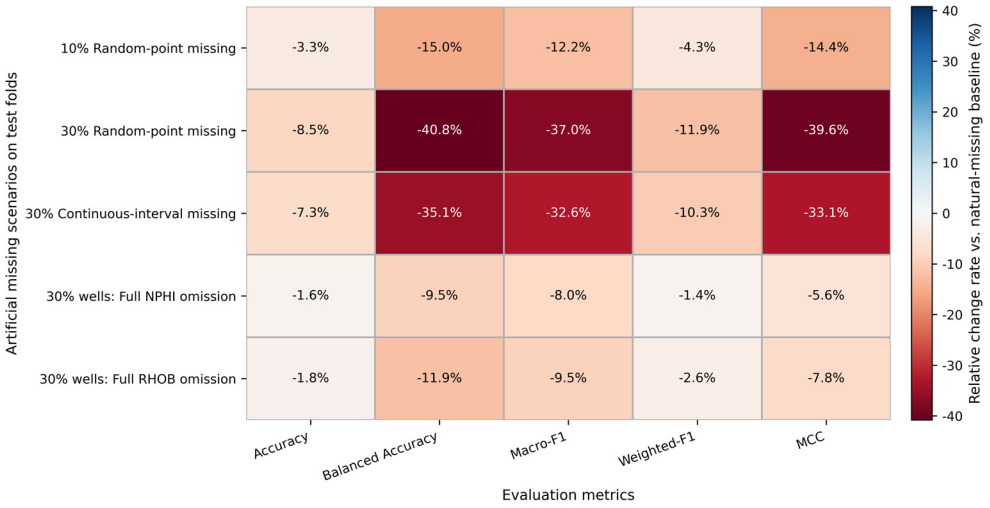

Fig 4. Performance changes under different synthetic missing scenarios relative to the natural missing baseline.

Random point missing produces a clear degradation trend. When the missing ratio increases from 0 to 10% and 30%, Accuracy decreases from 0.687 to 0.665 and 0.629, respectively, while Macro-F1 decreases from 0.328 to 0.288 and 0.207. At 30% random missing, Macro-F1 and MCC decline by 36.98% and 39.63% relative to the baseline.

Balanced Accuracy, Macro-F1, and MCC decrease more strongly than Accuracy and Weighted-F1, showing that missing logging data disproportionately affect minority-class discrimination. Shale remains comparatively robust because of its large sample size and broad response coverage, whereas minority lithologies rely on narrower feature combinations and are more easily absorbed by dominant classes when discriminative information is lost. Consequently, Accuracy alone may underestimate the impact of incomplete logging data.

At the nominal 30% setting, continuous-interval missing yields slightly higher Macro-F1 (0.221) and MCC (0.287) than random point missing (0.207 and 0.259). However, the two missing patterns affect different samples and curve combinations and therefore should not be interpreted as strictly equivalent experiments. Random point missing disperses information loss across a larger number of depth samples, whereas continuous missing is locally concentrated and its impact depends strongly on whether critical lithology intervals are affected.

When one entire curve is omitted from 30% of the test wells, model degradation is smaller in the aggregated results. Macro-F1 is 0.302 for NPHI omission and 0.297 for RHOB omission. However, entire-curve omission and random-point missing involve different amounts and

ISSN: 2688-8645

distributions of information loss and therefore cannot be directly compared under an equal missing-data budget.

RHOB omission produces slightly lower mean performance than NPHI omission, but paired Wilcoxon tests across the five folds yield $p>0.05p>0.05$ for all evaluation metrics. Thus, the results indicate only a tendency toward greater sensitivity to RHOB loss, rather than a statistically significant difference.

These results have three practical implications. First, logging quality control should consider not only the missing rate of individual curves but also the combinations of available curves at each depth. Second, tree models with native missing-value handling can still support preliminary lithology screening when a single curve is unavailable, but the missing status should be reported together with the prediction. Third, class-balanced metrics such as Balanced Accuracy, Macro-F1, and MCC should be used alongside Accuracy when evaluating models under incomplete logging conditions.

3.4. SHAP Feature Contributions and Geological Interpretation

Global SHAP importance under original-distribution and class-balanced sampling is compared in Fig. 5. Under original-distribution sampling, the mean absolute SHAP contribution shares are DTC 24.80%, RDEP 21.12%, GR 20.25%, RHOB 19.17%, and NPHI 14.65%. After class balancing, the ranking changes to RDEP (22.70%) > RHOB (21.67%) > DTC (20.89%) > GR (19.58%) > NPHI (15.16%).

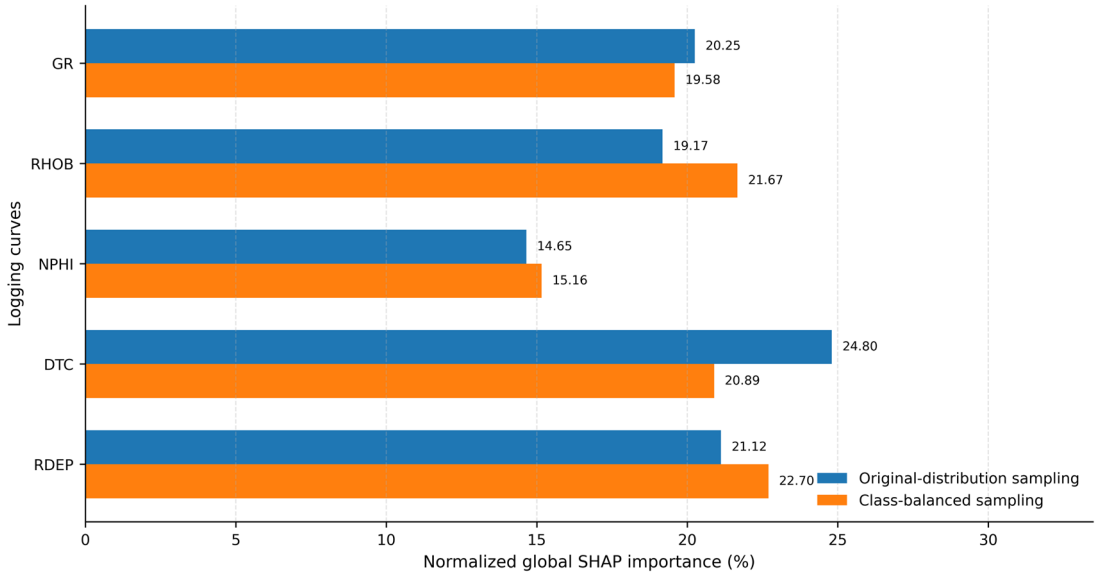

Fig 5. Comparison of global SHAP importance between original distribution sampling and class-balanced sampling.

The difference between the two rankings reflects the strong class imbalance in the dataset. Because shale accounts for 61.58% of all samples, original-distribution sampling gives greater explanatory weight to shale-related feature dependence. Class-balanced sampling assigns equal importance to each lithology and therefore increases the relative contribution of features important to minority classes, particularly RDEP and RHOB. The two sampling strategies consequently answer different questions: the former reflects average model behavior under the observed class distribution, whereas the latter reflects feature dependence when all lithologies are treated equally.

Across 2,000 stratified bootstrap resamples of the class-balanced explanation set, RDEP ranks first in 100% of iterations, while the remaining feature ranking is also largely stable. This result

ISSN: 2688-8645

demonstrates stability with respect to explanation-sample selection for the fixed model, but does not represent uncertainty associated with model retraining.

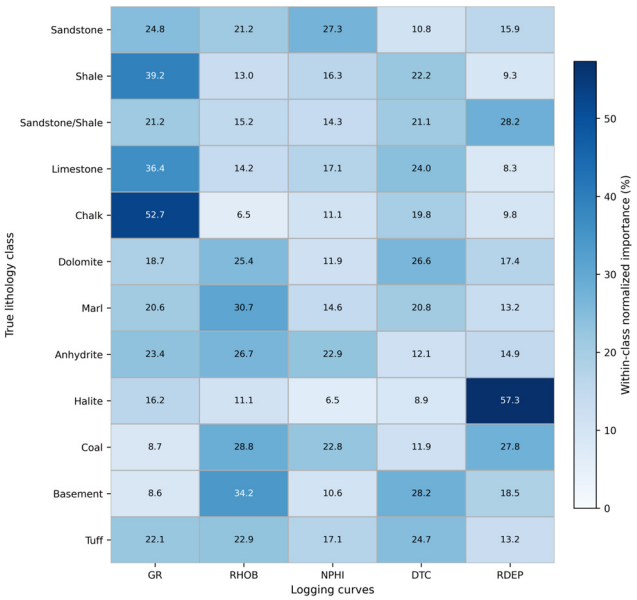

Fig 6. Normalized SHAP importance heatmap for different lithologies under class-balanced sampling

Class-wise SHAP results show that different lithologies depend on different logging curves. Shale is strongly influenced by GR, consistent with the radioactive response of argillaceous minerals. Sandstone is most influenced by NPHI in the fitted model, although this relationship should be interpreted as dataset-specific and multivariate rather than as a universal petrophysical rule. Sandstone/shale depends strongly on RDEP but remains difficult to classify, demonstrating that a high SHAP contribution does not necessarily imply high class separability. Limestone and chalk are sensitive to GR, anhydrite and coal show strong RHOB dependence, and halite depends mainly on RDEP.

The SHAP results are broadly consistent with the missing-data experiments. RHOB has greater class-balanced importance than NPHI, and removing RHOB also causes slightly greater average performance degradation. However, the two analyses measure different quantities: SHAP describes feature contributions under the existing input conditions, whereas omission experiments quantify performance loss after information is removed. Because logging curves contain correlated and redundant information, a higher SHAP value should not be interpreted as implying a proportional performance decline when that feature becomes unavailable.

4. Limitations and Outlook

This study has several limitations. First, the analysis is based solely on the FORCE 2020 dataset from the Norwegian Sea, and external validation in other basins is still required. Second, only five conventional logging curves are used. Mineral-sensitive logs, mud logs, core data, and stratigraphic information may improve the discrimination of difficult classes such as dolomite and chalk. Third, the current model performs point-wise classification and does not explicitly consider vertical continuity, while the synthetic missing-data experiments do not cover more complex quality problems such as curve drift, systematic offsets, and increased noise. Fourth, simple inverse-frequency weighting remains insufficient for highly imbalanced classes, and SHAP reflects model dependence rather than geological causality; its interpretation should therefore be treated as a diagnostic aid rather than a definitive geological criterion.

ISSN: 2688-8645

Future work should focus on richer multi-source features, sequence-aware or hierarchical models, more robust strategies for rare lithologies, and independent cross-basin validation combined with uncertainty calibration, with the aim of developing a practical workflow of "model prediction–risk alert–manual review."

5. Conclusion

Using the FORCE 2020 well-log dataset, this study develops a lithology-classification workflow integrating well-grouped cross-validation, missing-data robustness testing, OOF diagnosis, and TreeSHAP interpretation. The main conclusions are:

(1) XGBoost with native missing-value handling and no class weighting provides the best overall balance, achieving an Accuracy of 0.687, Macro-F1 of 0.328, and MCC of 0.430. Class weighting improves Balanced Accuracy but reduces overall performance.

(2) The model performs better for shale, anhydrite, sandstone, and limestone, but remains weak for transitional and rare classes. In particular, 67.02% of sandstone/shale samples are misclassified as shale, forming the dominant error path.

(3) Missing logs significantly reduce classification performance. Under 30% random point missing, Macro-F1 and MCC decrease by 36.98% and 39.63%, respectively. Different missing-data patterns affect different scopes and should not be directly compared under an assumed equal missing budget.

(4) Class-balanced SHAP identifies RDEP, RHOB, and DTC as the most influential features overall. The proposed workflow is suitable for cross-well lithology pre-screening and quality-control support under incomplete logging conditions, while ambiguous and rare lithologies still require manual review.

References

[1] Busch, J. M., Fortney, W. G., & Berry, L. N. (1987). Determination of lithology from well logs by statistical analysis. SPE Formation Evaluation, 2(4), 412–418.

[2] Zhu, X., Zhang, H., Ren, Q., et al. (2024). A review on intelligent recognition with logging data: Tasks, current status and challenges. Surveys in Geophysics, 45(5), 1493–1526. https://doi.org/ 10.1007 /s 10712-024-09862-9.

[3] Zou, W. B. (2020). Artificial intelligence research status and applications in well logging. Well Logging Technology, 44(4), 323–328.

[4] Wang, Z. Z., Zhang, C. L., & Gao, S. C. (2017). Identification of complex carbonate lithology using decision tree method: A case study of Sudong 41-33 block in Sulige Gas Field. Petroleum Geology and Recovery Efficiency, 24(6), 25–33.

[5] Liu, H., Wu, Y., Cao, Y., et al. (2020). Well logging based lithology identification model establishment under data drift: A transfer learning method. Sensors, 20(13), 3643. https:// doi.org/ 10. 3390/ s 20133643.

[6] Hall, B. (2016). Facies classification using machine learning. The Leading Edge, 35(10), 906–909. https://doi.org/10.1190/tle35100906.1.

[7] Halotel, J., Demyanov, V., & Gardiner, A. (2020). Value of geologically derived features in machine learning facies classification. Mathematical Geosciences, 52(1), 5–29. https://doi.org/ 10.1007/ s11004-019-09820-3.

[8] Dev, V. A., & Eden, M. R. (2019). Formation lithology classification using scalable gradient boosted decision trees. Computers & Chemical Engineering, 128, 392–404. https://doi.org/ 10. 1016/ j. compchemeng. 2019.06.014.

[9] Bressan, T. S., Kehl de Souza, M., Girelli, T. J., et al. (2020). Evaluation of machine learning methods for lithology classification using geophysical data. Computers & Geosciences, 139, 104475. https://doi.org/10.1016/j.cageo.2020.104475.

[10] Zhou, K., Zhang, J., Ren, Y., et al. (2020). A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification. Geophysics, 85(4), WA147–WA158. https://doi.org/10.1190/geo2019-0598.1.

[11] Merembayev, T., Kurmangaliyev, D., Bekbauov, B., et al. (2021). A comparison of machine learning algorithms in predicting lithofacies: Case studies from Norway and Kazakhstan. Energies, 14(7), 1896. https://doi.org/10.3390/en14071896.

[12] Zhu, L., Li, et al. (2018). Intelligent logging lithological interpretation with convolution neural networks. Petrophysics, 59(6), 799–810.

[13] Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering, 174, 216–228. https://doi.org/10.1016/j. petrol.2018.11.060.

[14] Liu, X., Shao, G., Liu, Y., et al. (2022). Deep classified autoencoder for lithofacies identification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14. https://doi.org/ 10.1109/ TGRS. 2022. 3141479.

[15] Feng, R. (2021). Uncertainty analysis in well log classification by Bayesian long short-term memory networks. Journal of Petroleum Science and Engineering, 205, 108816. https://doi. org/ 10. 1016/ j. petrol.2021.108816.

[16] Zeng, L., Ren, W., & Shan, L. (2020). Attention-based bidirectional gated recurrent unit neural networks for well logs prediction and lithology identification. Neurocomputing, 414, 153–171. https:// doi.org/10.1016/j.neucom.2020.07.060.

[17] Santos, D. T. dos, Roisenberg, M., & Nascimento, M. dos S. (2022). Deep recurrent neural networks approach to sedimentary facies classification using well logs. IEEE Geoscience and Remote Sensing Letters, 19. https://doi.org/10.1109/LGRS.2021.3131757.

[18] Bormann, P., Aursand, P., Dilib, F., et al. (2020). FORCE 2020 well log and lithofacies dataset for machine learning competition. Zenodo. https://zenodo.org/record/4351155.

[19] Ellis, D. V., & Singer, J. M. (2007). Well logging for earth scientists. Springer Netherlands. https:// doi. org/10.1007/978-1-4020-5514-6.

[20] Al-Mudhafar, W. J. (2017). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. Journal of Petroleum Exploration and Production Technology, 7(4), 1023–1033. https://doi.org/10.1007/s13202-017-0368-1.

[21] Feng, R. (2021). Uncertainty analysis in well log classification by Bayesian long short-term memory networks. Journal of Petroleum Science and Engineering, 205, 108816. https://doi.org/10.1016/j. petrol. 2021.108816.

[22] Zhang, J., He, Y., Zhang, Y., et al. (2022). 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/ en 15 103675.

[23] Sun, Y., Pang, S., Zhao, Z., et al. (2024). Interpretable SHAP model combining meta-learning and vision transformer for lithology classification using limited and unbalanced drilling data in well logging. Natural Resources Research, 33(6), 2545–2565. https://doi.org/10.1007/s11053-024-09724-1.

[24] Li, N., Xu, B. S., Wu, H. L., Feng, Z., Li, Y. S., Wang, K. W., Liu, P. (2021). Application status and prospects of artificial intelligence in well logging and formation evaluation. Acta Petrolei Sinica, 42(4), 508–522.

[25] Carvalho, H. M. O., Daniel, H., Korenchendler, A., et al. (2026). Lithology classification based on well log data: A benchmark for machine learning models. Mathematical Geosciences.

[26] Jiang, C., Zhang, D., & Chen, S. (2021). Lithology identification from well-log curves via neural networks with additional geologic constraint. Geophysics, 86(5), IM85–IM100. https://doi.org/ 10. 1190/ geo2020-0472.1.

[27] Xie, D., Liu, Z., Wang, F., et al. (2024). A transformer and LSTM-based approach for blind well lithology prediction. Symmetry, 16(5), 616. https://doi.org/10.3390/sym16050616.

[28] Wang, J., & Cao, J. (2023). A lithology identification approach using well logs data and convolutional long short-term memory networks. IEEE Geoscience and Remote Sensing Letters, 20, 1–5. https: // doi. org/10.1109/LGRS.2023.3241318.

[29] Cao, Z. M., Liu, P. C., Han, J., et al. (2025). A lithology identification method based on meta-learning using multi-view resampled heterogeneous representative features from well logs. Geophysical Prospecting for Petroleum, 64(3), 575–586.

[30] Shao, X., Zhang, P., Yan, S., et al. (2025). A multi-model fusion network for enhanced blind well lithology prediction. Processes, 13(1), 278. https://doi.org/10.3390/pr13010278.

[31] Carvalho, H. M. O., Daniel, H., Korenchendler, A., et al. (2026). Lithology classification based on well log data: A benchmark for machine learning models. Mathematical Geosciences.

[32] Dong, S., Wang, L., Xu, T., et al. (2026). Improved recurrent neural network for complex lithology identification in gas reservoirs. Petroleum Science and Technology, 44(16), 2673–2694. https:// doi. org/10.1080/10916466.2026.2462118.

[33] Gama, P. H. T., Faria, J., Sena, J., et al. (2025). Imputation in well log data: A benchmark for machine learning methods. Computers & Geosciences, 196, 105789. https://doi.org/10. 1016/j. cageo. 2025. 105789.

[34] Garini, S. A., Shiddiqi, A. M., & Utama, W. (2025). Filling-well: An effective technique to handle incomplete well-log data for lithology classification using machine learning algorithms. MethodsX, 14, 103127. https://doi.org/10.1016/j.mex.2025.103127.

[35] Pang, Q., Chen, C., Li, W., et al. (2026). Lithology identification from missing well log data: a multi-constraint guided self-supervised diffusion framework. Advanced Engineering Informatics, 74, 104697. https://doi.org/10.1016/j.aei.2026.104697.

[36] Zhu, X., Zhang, H., Ren, Q., et al. (2023). An automatic identification method of imbalanced lithology based on Deep Forest and K-means SMOTE. Geoenergy Science and Engineering, 224, 211595. https: // doi.org/10.1016/j.geoen.2023.211595.

[37] IEEE. (2026). Causally informed domain-guided machine learning for imbalanced lithofacies classification from well logs. IEEE Xplore. [2026-07-30].

[38] Ekmekcioğlu, Ö, Koc, K., Özger, M., et al. (2022). Exploring the additional value of class imbalance distributions on interpretable flash flood susceptibility prediction in the Black Warrior River basin, Alabama, United States. Journal of Hydrology, 610, 127877. https://doi.org/10. 1016/j. jhydrol. 2022. 127877.

[39] Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874.

[40] Fetni, Y., Ameur-Zaimeche, O., Wood, D. A., et al. (2025). Explainable machine-learning workflow to improve real-time lithofacies identification using mudlogging data from a heterogeneous clastic reservoir sequence. Computational Geosciences, 29(6), 56. https://doi.org/10.1007/s10596-025-10367-1.

[41] Sun, Y., Pang, S., Li, H., et al. (2025). Enhanced lithology classification using an interpretable SHAP model integrating semi-supervised contrastive learning and transformer with well logging data. Natural Resources Research, 34(2), 785–813. https://doi.org/10.1007/s11053-025-09916-9.