Balanced Learning with Optimized Extra Trees Classifier for Reliable Lithology Identification in Imbalanced Well Log Data

Ali Daneshpour
Dep. of Math. & Computer Science
Amirkabir University of Technology
(Tehran Polytechnic), Iran
alidaneshpour@aut.ac.ir

Behnam Yousefimehr
Dep. of Math. & Computer Science
Amirkabir University of Technology
(Tehran Polytechnic), Iran
behnam.y2010@aut.ac.ir

Mehdi Ghatee
Dep. of Math. & Computer Science
Amirkabir University of Technology
(Tehran Polytechnic), Iran
ghatee@aut.ac.ir
Corresponding author

***Abstract*—Accurate lithology identification is crucial for subsurface characterization in hydrocarbon exploration, yet conventional methods often fail to capture the complex, nonlinear relationships in well log data. To address this challenge, we propose a robust machine learning framework based on an optimized Extra Trees Classifier, enhanced by a hybrid resampling strategy to mitigate severe class imbalance. Our approach combines random oversampling of minority lithologies (e.g., coal and dolomite) with strategic undersampling of dominant classes, ensuring balanced representation while preserving critical geological patterns. Hyperparameter tuning via optimization further refines model performance, achieving an accuracy of 83.91% with a penalty score of -0.4087, demonstrating superior reliability, particularly for underrepresented facies. A comparative computational analysis confirms the efficiency of our framework, which outperforms complex models, such as GrowNet and Blender, as well as baselines, in both speed and scalability. To promote reproducibility, we provide the complete implementation, including preprocessing scripts and trained models, at https://github.com/alidaneshpour/ICCKE-2025.**

***Index Terms*—Lithology Identification, Extra Trees Classifier, Machine Learning, Well Logs, Imbalanced Data.**

I. Introduction

Accurate lithological identification is essential for oil and gas exploration, as well as for reservoir geological evaluation [1]. However, the inherently complex and often nonlinear relationships between well logging parameters and lithological properties present significant challenges for traditional interpretation methods [2]. In recent years, artificial intelligence (AI) techniques have gained increasing attention in this domain, enabling the use of advanced data mining algorithms that markedly enhance lithology classification accuracy based on well log data [3], [4].

Among these techniques, ensemble learning methods have shown particular promise due to their robustness and ability to model nonlinearities in high-dimensional datasets [5]. Nevertheless, a major bottleneck in lithology classification remains the class imbalance issue commonly observed in real-world well log data, where dominant lithofacies, such as shale and sandstone, vastly outnumber underrepresented classes, including coal, dolomite, and limestone. This imbalance often leads to biased models that perform poorly on minority classes, limiting the practical applicability of such classifiers [6].

In this study, we propose an effective and computationally efficient framework for lithology identification using the Extra Trees Classifier, a randomized ensemble learning algorithm known for its strong generalization performance and low computational overhead. We address the class imbalance problem through a hybrid resampling strategy combining random oversampling and undersampling techniques, ensuring improved sensitivity to minority classes while maintaining overall classification integrity. Additionally, we leverage the AutoGluon [7] framework for hyperparameter tuning to optimize model performance across lithofacies categories.

The novelty of our work lies in four key contributions:

- Achieving high classification performance on a challenging, imbalanced lithology dataset;
- Maintaining excellent computational efficiency suitable for large-scale deployment;
- Effectively handling class imbalance using a simple yet powerful hybrid resampling strategy;
- Demonstrating robustness of the model through noise analysis and sensitivity evaluation.

Through a detailed performance evaluation and time complexity analysis, we demonstrate that our approach strikes a practical balance between accuracy, scalability, and robustness, offering a deployable solution for real-world subsurface characterization.

The structure of the paper is as follows: Section II reviews related work, Section III details our methodology, Section V presents the results and evaluation, and Section VII concludes the paper and outlines future research directions.

II. Related Works

This section provides a comprehensive review of the most significant studies on lithology prediction employing artificial intelligence techniques. As an example, [8] utilized random forest and deep learning models (CNN and ResNet) to predict

979-8-3315-8973-8/25/$31.00 © 2025 IEEE

lithofacies based on gamma ray, density, and resistivity logs from 22 wells in Southern Sichuan, China. Furthermore, [9] introduces an advanced lithology classification framework that was developed by integrating physics-based feature engineering, SMOTE-NearMiss balancing, and novel machine and deep learning models, including GrowNet, Deep-Insight, and a model blender.

Well logs are essential tools in subsurface exploration. however, they frequently contain missing values due to various operational challenges. This issue can hinder the practical use of well logs in real-world applications, such as lithology identification. This challenge is further complicated by the varying presence and distribution of missing logs across different wells and depth intervals. A recent work [10] introduces FlexLogNet, a deep learning framework specifically designed for well-log completion. This approach leverages a hybrid architecture that combines a heterogeneous graph neural network with a fully connected network. This integration enables the model to predict various missing well log types flexibly, offering a promising solution for handling incomplete well log data. Another study [11] introduces a multivariate well log imputation method using the MICE framework with machine learning regressors. The study demonstrates that iterative, chained prediction with Gradient Boosted Trees enables robust reconstruction of missing logs in complex, sparse datasets. This approach outperforms traditional single-target methods.

The inherent heterogeneity of the field's stratigraphic layers leads to a natural imbalance in lithology datasets. Numerous studies have sought to mitigate the degradation in predictive performance caused by class imbalance in machine learning models. For instance, [12] addresses data imbalance in lithology identification by enhancing the representation of thin layers and minority classes using Max-min-distance PCA correction. Additionally, an attention-augmented graph convolutional network is employed, trained on both labeled and unlabeled samples. Also, [13] addresses the multiclass imbalance problem by employing error-correcting output code (ECOC)-based decomposition and cost-sensitive learning within a heterogeneous ensemble. This approach transforms the original task into a series of balanced binary subproblems, effectively mitigating misclassifications of the minority class. Frontier approaches emphasize enhancing model robustness by addressing the challenge of noisy labels [14], [15].

III. Methodology

This section presents a detailed overview of the methodology employed in this study, outlining the techniques and approaches utilized. Before introducing the proposed framework, we first describe the dataset used in this research. An illustration of the overall methodology is provided in Figure 1.

A. Data Preprocessing

Comprehensive data preprocessing is essential for successful lithofacies classification. Initially, the provided well log datasets were imported as dedicated training and test sets. Column inspection was performed to identify irrelevant or redundant features; some column was removed from both splits to streamline the feature set and focus on predictive attributes essential for lithofacies prediction.

Subsequently, a subset of instances was randomly sampled from the training data to manage computational requirements and enable efficient model development. All categorical labels were preserved in their native format to ensure compatibility with downstream machine learning algorithms. This preprocessing phase established a consistent, well-structured dataset for subsequent analysis.

B. Extra Trees Classifier

The Extra Trees Classifier, short for Extremely Randomized Trees [16], is an ensemble-based machine learning algorithm that constructs a large number of unpruned decision trees and aggregates their outputs to form a robust prediction. It is beneficial for classification tasks involving high-dimensional and noisy data, such as well logs used for lithology identification.

Unlike conventional decision tree algorithms that deterministically select the best feature and threshold based on a specific impurity criterion, Extra Trees introduces randomness at both the feature selection and threshold determination stages. More specifically, the algorithm selects a subset of candidate features at each split and then draws thresholds for splitting these features randomly, rather than optimizing over all possible values. This randomness helps increase model diversity and typically results in improved generalization performance when predictions from all trees are averaged [17].

Formally, the predicted class label $\hat{y}$ for an input sample $\mathbf{x}$ is obtained by majority voting over the predictions of T individual randomized decision trees:

$$\hat{y} = \text{mode}\left(\{h_t(\mathbf{x})\}_{t=1}^{T}\right), \tag{1}$$

where $h_t(\mathbf{x})$ represents the prediction of the t-th decision tree [16].

Each tree is built recursively by splitting nodes based on randomly selected features and thresholds. At each internal node, K features are chosen at random from the feature set $\mathcal{F}$, and for each selected feature f_k, a threshold θ_k is randomly drawn from the range of that feature's values in the current node. The best split is chosen based on an impurity reduction criterion such as the Gini index:

$$G(S) = 1 - \sum_{c=1}^{C} p_c^2, \tag{2}$$

where p_c denotes the proportion of samples belonging to class c within node S, and C is the total number of classes .

For each candidate split (f_k, θ_k), the impurity reduction is computed using:

$$\Delta G = G(S) - \left(\frac{|S_L|}{|S|}G(S_L) + \frac{|S_R|}{|S|}G(S_R)\right), \tag{3}$$

where S_L and S_R are the left and right child nodes resulting from the split. The split yielding the maximum ΔG is selected [16].

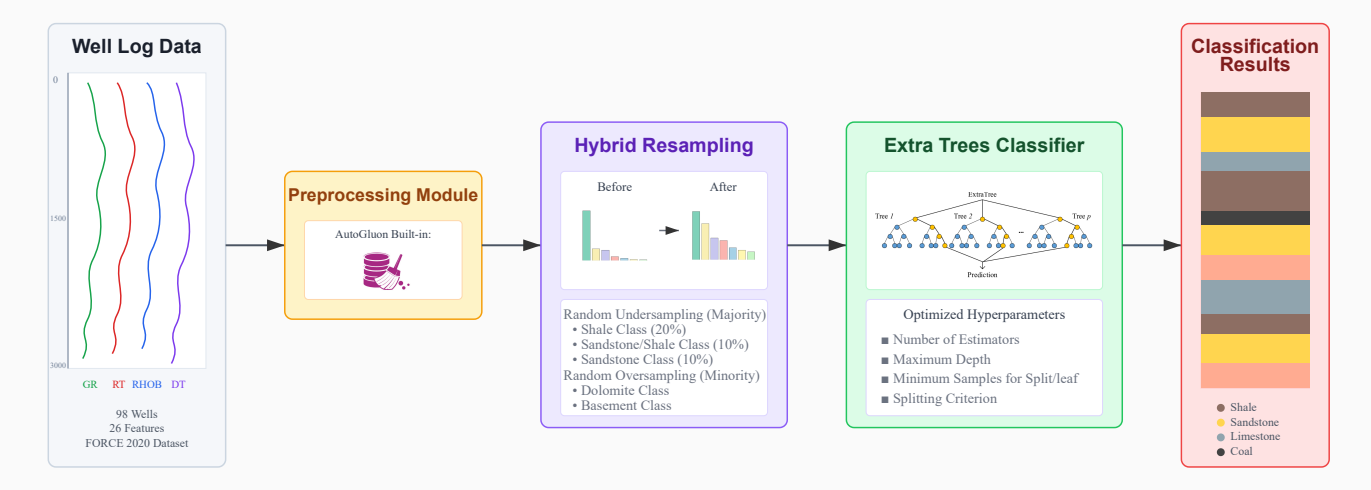

Fig. 1. Overview of the proposed methodology.

This approach is particularly well-suited for lithology classification tasks. The stochastic nature of the Extra Trees algorithm helps mitigate overfitting and enhances its ability to capture the complex, nonlinear relationships inherent in well log data. Furthermore, because Extra Trees does not rely on bootstrapped samples and uses highly randomized splits, it often performs well in the presence of class imbalance, common in geological datasets, by reducing the model's sensitivity to dominant classes and over-represented patterns [18].

In this study, the Extra Trees Classifier is employed to learn discriminative patterns from multi-dimensional well log features and classify subsurface lithologies. Its ability to handle high variance, model nonlinearity, and resist overfitting makes it an effective tool for the lithological identification problem on imbalanced datasets.

IV. Resampling Strategy

Class imbalance is a significant challenge in machine learning and lithology classification tasks, especially when working with real-world well log datasets [19], [20]. As illustrated in Figure 2, the dataset used in this study is highly imbalanced, with a disproportionately large number of samples belonging to the *Shale* and *Sandstone* classes. These lithologies dominate the dataset, while other classes such as *Limestone*, *Dolomite*, and *Coal* are underrepresented, forming the minority classes.

To address this imbalance, we applied a hybrid resampling approach combining random undersampling of the majority classes and random oversampling of the minority classes. The aim was to create a more balanced class distribution, which is essential for improving classifier sensitivity to underrepresented lithologies and mitigating bias toward the majority classes.

In random undersampling, a subset of samples from the majority classes (Shale and Sandstone) is selected at random and removed from the training set. This reduces the dominance of these classes and helps prevent the model from overfitting to them. Simultaneously, random oversampling is performed by duplicating samples from minority classes until their representation is comparable to that of the majority classes. While this may introduce some redundancy, it allows the classifier to better learn the decision boundaries associated with minority lithologies.

We also evaluated more sophisticated sampling techniques, including Synthetic Minority Over-sampling Technique (SMOTE) [21], Adaptive Synthetic Sampling (ADASYN) [22], and ClusterCentroids [23]. Although these methods can generate synthetic samples to better represent minority classes, we found that they significantly increased training time and added complexity without consistently improving classification performance in our use case.

Considering the trade-off between computational efficiency and predictive performance, we opted to use random over- and undersampling methods. This approach provided a balanced and practical solution for our workflow, effectively enhancing the classifier's performance on minority lithology classes without introducing substantial overhead.

A. Hyperparameter Tuning

Effective hyperparameter tuning plays a critical role in optimizing the performance of machine learning models, particularly for high-dimensional, imbalanced datasets such as those encountered in lithology classification from well logs. In this study, we leveraged the capabilities of the `AutoGluon` framework to implement and fine-tune the Extra Trees Classifier.

To ensure reliable and reproducible results, a balanced training subset was created after applying custom sampling strategies to reduce overrepresentation of certain lithofacies classes.

Hyperparameter optimization was carried out using 5-fold cross-validation across a wide search space. The main parameters considered were:

- **Number of estimators** (100–1000): Larger ensembles generally reduce variance and improve generalization.
- **Maximum depth** (50–250): Regulates model complexity and helps capture non-linear patterns.
- **Minimum samples for split/leaf** (2–15): Prevents overfitting by enforcing thresholds for splitting and leaf creation.
- **Splitting criterion** (Gini): Determines how split quality is evaluated.

V. Results And Evaluation

This section presents a comprehensive evaluation of the proposed lithofacies classification methodology, including quantitative performance metrics and error analysis. The evaluation metrics, and comparative analysis with existing approaches are described in detail.

To promote transparency and facilitate reproducibility, the complete source code, along with all preprocessing scripts and trained model checkpoints, is publicly available on GitHub[1].

A. Dataset Description

The FORCE 2020 lithology prediction dataset [24] is a publicly available benchmark compiled through a collaboration between Norwegian governmental authorities and the petroleum industry. It comprises well-log and interpreter-labeled lithofacies data from 98 offshore wells located in the South Viking Graben and North Viking Graben regions of the North Sea. The dataset has a sampling resolution of 0.15 meters and includes 26 petrophysical and drilling-related features, such as resistivity, sonic, gamma ray, porosity, and caliper logs, in addition to spatial coordinates and depth information.

Twelve lithofacies classes are defined in the dataset, each accompanied by interpreter-assigned confidence levels. As illustrated in Figure 2, the lithofacies distribution is highly imbalanced, posing additional challenges for effective model training. A notable issue associated with this dataset is the prevalence of missing values across many of the logging measurements and depth intervals, with only gamma ray and depth features consistently available for all samples. This characteristic necessitates the use of robust imputation strategies or the development of model architectures capable of handling incomplete data inputs.

For model evaluation, the dataset provides both open and closed blind test sets, each comprising 10 wells. The combination of diverse geological settings, realistic data quality, and inherent noise makes this dataset particularly valuable for advancing machine learning research in subsurface characterization and lithofacies classification. The FORCE 2020 dataset used in this study is publicly available on GitHub[2].

[1]https://github.com/alidaneshpour/ICCKE-2025

[2]https://github.com/bolgebrygg/Force-2020-Machine-Learning-competition

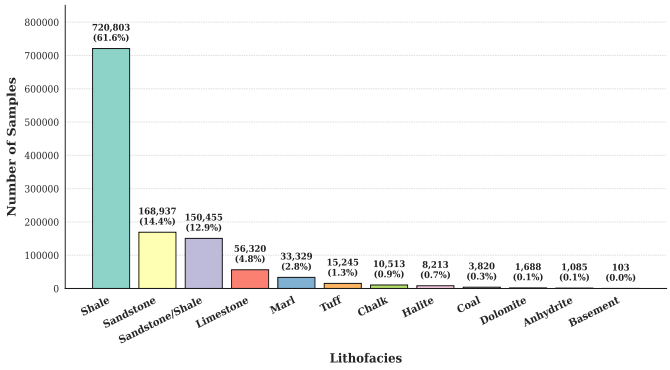

Fig. 2. Distribution of Lithofacies in Dataset

B. Performance Assessment Criteria

To thoroughly assess the model's performance, we employed the following evaluation criteria:

- **Accuracy:** Calculated as the proportion of correctly classified samples to the total number of samples. While accuracy provides an overall indication of performance, it does not distinguish between the types of misclassification or address class imbalance.
- **Penalty Score:** A geologically-informed metric that quantifies the severity of misclassifications using a penalty matrix (see Figure 9 in [9]). For example, predicting Marl instead of Shale incurs a smaller penalty than predicting Anhydrite instead of Shale, as the former error is more geologically plausible. The penalty score is computed as:

$$P = -\frac{1}{N}\sum_{i=1}^{N} A_{y_i,\hat{y}_i} \tag{4}$$

  where N is the number of samples, y_i is the predicted class for the i-th sample, $\hat{y}_i$ is the true class, and A represents the penalty matrix [9]. A penalty score closer to zero indicates more geologically plausible classifications, while larger negative values reflect more severe and geologically implausible misclassifications.
- **Confusion Matrix:** Offers a detailed view of the distribution of predicted versus true classes, providing insights into which lithofacies are most frequently confused by the model.

C. Experimental Results

The performance of the proposed lithofacies classification model is summarized in this section. Figure 3 presents the normalized confusion matrix, which visualizes the distribution of predicted versus actual lithofacies classes. This allows for a quick assessment of both correct and incorrect classification patterns across all classes.

A quantitative comparison with previous studies and baselines is given in Table I and II, reporting both accuracy and penalty score for various algorithms. The Extra Trees Classifier

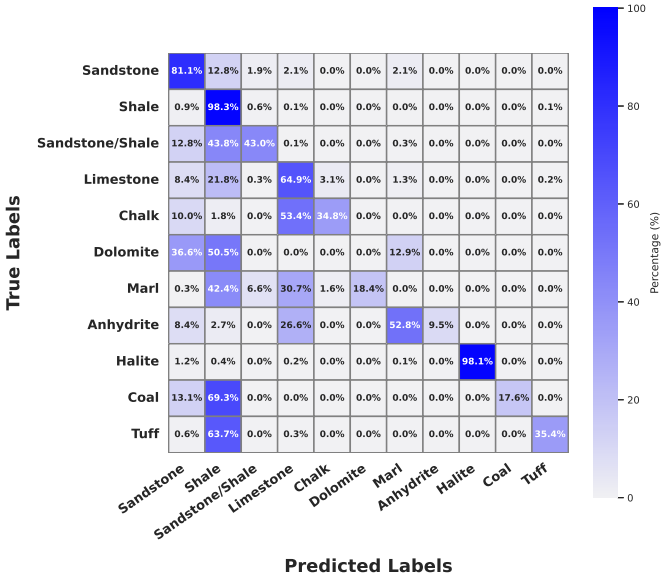

Fig. 3. Normalized confusion matrix illustrating the distribution of model predictions across actual lithofacies classes.

introduced in this study achieves the highest accuracy and the lowest penalty score among all compared methods, indicating improved classification performance.

TABLE I
COMPARISON OF THIS STUDY'S RESULTS WITH BASELINES.

Model	Accuracy	precision	Recall	F1 Score
Decision Tree	62.37	44.31	41.74	40.78
Random Forest	74.20	64.29	46.10	49.96
SVM	58.68	5.33	9.09	6.72
KNN	55.91	30.49	30.94	28.96
Logistic Regression	58.68	5.33	9.09	6.72
Our: Extra Trees	**83.91**	**79.93**	**49.48**	**60.74**

TABLE II
COMPARISON OF THIS STUDY'S RESULTS WITH PREVIOUS STUDIES, IN TERMS OF ACCURACY AND PENALTY SCORE.

Study	Final Model	Accuracy	Penalty Score
[9]	Blender	83	-0.4303
[9]	GrowNet	83	-0.4145
This Study	**Extra Trees**	**83.91**	**-0.4087**

D. Time Complexity Analysis of the Extra Trees Classifier

In theoretical terms, the Extra Trees Classifier trains in $\mathcal{O}(T\,N\log N)$ time, where T is the number of trees and N the number of training samples. At each node, split thresholds are drawn at random, avoiding the $\mathcal{O}(N\log N)$ sorting step per feature that a classical Random Forest requires; this reduces the constant factors in practice while retaining the same asymptotic bound. Inference on a single sample then costs $\mathcal{O}(T\log N)$, since traversal down each tree halts at depth proportional to $\log N$.

By comparison, a nonlinear Support Vector Machine can exhibit training complexity between $\mathcal{O}(N^2P)$ and $\mathcal{O}(N^3)$ (with P features), making it impractical for large N. The k-Nearest Neighbors classifier defers all work to prediction time, incurring $\mathcal{O}(N\,P)$ per query. Gradient-boosted ensembles such as XGBoost or CatBoost formally share the $\mathcal{O}(M\,N\log N)$ bound (with M boosting rounds), but each round must compute gradients and Hessians and is inherently sequential, inflating constant factors and end-to-end runtime. Deep architectures (MLP, CNN) typically require $\mathcal{O}(E\,N\,P\,F)$ operations, where E is the number of epochs and F reflects network parameters and filter sizes, often demanding orders of magnitude more time unless specialized hardware is available.

Thus, the Extra Trees Classifier achieves the best of both worlds: it matches or exceeds the predictive performance of more complex models (see Table II) while maintaining an efficient $\mathcal{O}(T\,N\log N)$ training profile, low-overhead splits, and near-linear parallel scaling. These properties make it ideally suited for large-scale and time-sensitive lithofacies classification tasks.

VI. NOISE SENSITIVITY ANALYSIS

To assess the robustness of our lithofacies classification model, we conducted a noise sensitivity analysis by incrementally adding artificial noise to both the feature set and the labels in the test data. Gaussian noise was applied to the features, and random label flipping was used to corrupt the labels, both individually and in combination, at noise levels ranging from 5% to 30%.

The results, presented in Table III, demonstrate that the model retains high accuracy under increasing feature noise, with only a modest decrease from 0.83 to 0.78 as noise grows from 0% to 30%. This insensitivity to feature perturbations indicates that the proposed model is robust to moderate levels of measurement errors or noise in the input features. In contrast, the introduction of label noise leads to a more pronounced drop in classification accuracy, confirming that accurate labeling is critical for optimal model performance.

TABLE III
MODEL ACCURACY UNDER INCREASING LEVELS OF FEATURE NOISE, LABEL NOISE, AND BOTH.

Noise Level	**Feature Noise**	**Label Noise**	**Both**
0%	0.83	0.83	0.83
5%	0.82	0.78	0.78
10%	0.81	0.74	0.73
15%	0.81	0.70	0.69
20%	0.80	0.66	0.64
25%	0.79	0.62	0.60
30%	0.78	0.58	0.55

In summary, our model demonstrates strong robustness to feature noise, making it well-suited for deployment in real-world settings where input data may be subject to moderate levels of uncertainty or noise.

VII. Conclusion and Future Works

This study has demonstrated that the Extra Trees Classifier, when combined with a hybrid resampling strategy, provides a robust and computationally efficient solution for lithology identification in the context of highly imbalanced and incomplete well log datasets. The proposed approach achieved state-of-the-art performance on the FORCE 2020 benchmark, with an accuracy of 83.91% and a notably low penalty score of -0.4087, outperforming existing methods. These achievements underscore the substantial potential of advanced ensemble techniques and thoughtfully designed sampling strategies for tackling the inherent complexities of real-world subsurface data.

Despite these promising results, several avenues remain open for further research. A persistent challenge in real-world well log interpretation is the frequent occurrence of missing data, often due to the harsh environments encountered during drilling operations. Future work could explore the use of advanced data imputation techniques, such as Physics-Informed Neural Networks (PINNs), which explicitly model the physical relationships among logging parameters to enhance the reliability and robustness of classification. Additionally, employing knowledge distillation strategies may facilitate the deployment of more computationally demanding deep learning models by transferring their predictive capabilities to smaller, faster surrogate models, thereby addressing time complexity constraints in practical applications.

Overall, the integration of sophisticated imputation methods and efficient model compression techniques represents a promising direction for the advancement of automated lithology identification in increasingly complex and realistic subsurface scenarios.

References

[1] C. M. Saporetti, L. G. da Fonseca, E. Pereira, A lithology identification approach based on machine learning with evolutionary parameter tuning, IEEE Geoscience and Remote Sensing Letters 16 (12) (2019) 1819–1823.

[2] X. Ren, J. Hou, S. Song, Y. Liu, D. Chen, X. Wang, L. Dou, Lithology identification using well logs: A method by integrating artificial neural networks and sedimentary patterns, Journal of Petroleum Science and Engineering 182 (2019) 106336.

[3] X. Zhu, H. Zhang, Q. Ren, D. Zhang, F. Zeng, X. Zhu, L. Zhang, An automatic identification method of imbalanced lithology based on deep forest and k-means smote, Geoenergy Science and Engineering 224 (2023) 211595.

[4] H. Qian, Y. Geng, H. Wang, Lithology identification based on ramified structure model using generative adversarial network for imbalanced data, Geoenergy Science and Engineering 240 (2024) 213036.

[5] B. Yousefimehr, M. Ghatee, A. Heydari, Improving adhd detection with cost-sensitive lightgbm, in: 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, 2024, pp. 109–113.

[6] S. Yin, X. Lin, Z. Zhang, X. Li, A class-rebalancing self-training semisupervised learning for imbalanced data lithology identification, Geophysics 89 (1) (2024) WA1–WA11.

[7] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, arXiv preprint arXiv:2003.06505 (2020).

[8] Y. Shi, J. Liao, L. Gan, R. Tang, Lithofacies prediction from well log data based on deep learning: A case study from southern sichuan, china, Applied Sciences 14 (18) (2024).

[9] S. H. Mousavi, S. Hosseini-Nasab, A novel approach to classify lithology of reservoir formations using grownet and deep-insight with physic-based feature augmentation, Energy Science & Engineering 12 (2024) 4453–4477.

[10] C. Dai, X. Si, X. Wu, FlexLogNet: A flexible deep learning-based well-log completion method of adaptively using what you have to predict what you are missing, Computers and Geosciences 191 (2024) 105666.

[11] A. Hallam, D. Mukherjee, R. Chassagne, Multivariate imputation via chained equations for elastic well log imputation and prediction, Applied Computing and Geosciences 14 (2022) 100083.

[12] A. Wang, S. Zhao, K. Xie, C. Wen, H.-l. Tian, J.-B. He, W. Zhang, Attention mechanism-enhanced graph convolutional neural network for unbalanced lithology identification, Scientific Reports 14 (07 2024).

[13] M. S. Jamshidi Gohari, M. Emami Niri, S. Sadeghnejad, J. Ghiasi-Freez, An ensemble-based machine learning solution for imbalanced multiclass dataset during lithology log generation, Scientific Reports 13 (12 2023).

[14] X. Zhu, H. Zhang, R. Zhu, Q. Ren, L. Zhang, Classification with noisy labels through tree-based models and semi-supervised learning: A case study of lithology identification, Expert Systems with Applications 240 (2024) 122506.

[15] X. Feng, H. Luo, C. Wang, H. Gu, Reducing the effect of incorrect lithology labels on the training of deep neural networks for lithology identification, Mathematical Geosciences 56 (09 2023).

[16] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine learning 63 (1) (2006) 3–42.

[17] B. S. Bhati, C. Rai, Ensemble based approach for intrusion detection using extra tree classifier, in: Intelligent Computing in Engineering: Select Proceedings of RICE 2019, Springer, 2020, pp. 213–220.

[18] V. K. Naramala, L. P. Ravipudi, S. P. R. Bhavanam, V. N. Pokuri, V. K. Kishore, et al., Prediction of credit card fraud detection using extra tree classifier and data balancing methods, in: 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC), IEEE, 2024, pp. 722–728.

[19] B. Yousefimehr, M. Ghatee, A distribution-preserving method for resampling combined with lightgbm-lstm for sequence-wise fraud detection in credit card transactions, Expert Systems with Applications 262 (2025) 125661.

[20] B. Yousefimehr, M. Ghatee, M. A. Seifi, J. Fazli, S. Tavakoli, Z. Rafei, S. Ghaffari, A. Nikahd, M. R. Gandomani, A. Orouji, et al., Data balancing strategies: A survey of resampling and augmentation methods, arXiv preprint arXiv:2505.13518 (2025).

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[22] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), IEEE, 2008, pp. 1322–1328.

[23] S.-Y. R. Yen, Y.-C. Lee, Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset, in: Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on, Vol. 2, IEEE, 2006, pp. 4546–4551.

[24] P. Bormann, P. Aursand, F. Dilib, FORCE Machine Learning Competition, https://github.com/bolgebrygg/Force-2020-Machine-Learning-competition (Nov. 2020).