Bulletin of Geophysics and Oceanography Vol. XX, n. X, pp. XXX-XXX; Xxxxx 20xx
DOI 10.4430/bgo00540

Optimising sliding parameters in entropy-based well-log analysis for lithology discrimination using machine learning and a novel informational richness indicator

F.A. Cheddad, S. Makhlouf, A. Boulassel and Z. Boumelit

Geophysics Department – University of Sciences and Technology Houari Boumediene, Algiers, Algeria

(Received: 6 May 2025; accepted: 31 March 2026; published online: 19 October 2026)

ABSTRACT Entropy-based descriptors extracted from geophysical well logs are effective for lithology discrimination, but their performance strongly depends on sliding-window design, defined by window size (*W*) and step size (*S*). This study proposes a data-richness-driven workflow that combines four entropy measures, i.e. approximate entropy, permutation entropy, singular value decomposition entropy, and fuzzy entropy, with multiple machine learning models, including random forest (RF), gradient boosting (GB), multilayer perceptron (MLP), K-nearest neighbours, and decision tree. We introduce the informational richness indicator (IRI) and its normalised form (IRI%) to quantify the effective information density induced by *W* and *S*. Using a large, heterogeneous well-log dataset, results show that accuracy, precision, recall, and the F1-score increase monotonically with IRI% and stabilise beyond a high-information threshold. Three regimes are identified: low IRI% regime (<5%) with high variance and poor performance, mid IRI% regime (5–50%) with rapid improvement, and high IRI% regime (>50%) where performance plateaus. RF achieves the most robust results (≈93–96%), followed by MLP and GB. Targeting IRI% ≥ 50%, typically achieved with *W* = 60–100 and *S* ≤ 3, provides a practical and transferable rule for optimising entropy-based lithology discrimination while balancing accuracy and computational cost.

Key words: entropy-based profiling, geophysical well logs, sliding window optimisation, window size, step size, informational richness indicator, lithology discrimination, random forest, gradient boosting, multilayer perceptron, K-nearest neighbours, decision tree.

1. Introduction

Geophysical well logs are fundamental tools for the detailed characterisation of petroleum reservoirs, widely applied across geophysical and geological disciplines. They offer continuous, high-resolution subsurface data that are critical for evaluating lithological and petrophysical properties such as rock type, porosity, permeability, and fluid saturation (Bosch *et al.*, 2002; Dewhurst, 2015). Compared to core samples, which are limited to specific intervals, and seismic methods, which often lack the resolution to detect fine-scale variations, well logs provide a more consistent and comprehensive view of the reservoir (Gilbert *et al.*, 2004; Mikes *et al.*, 2006; Ren *et al.*, 2019; Imran *et al.*, 2020; Sun *et al.*, 2020).

© 2026 - The Author(s)

In lithology discrimination, well logs are particularly valuable, as they record a range of physical properties including acoustic velocity, resistivity, natural gamma radiation, and bulk density (RHOB), along the entire wellbore. These measurements enable geoscientists to delineate lithological boundaries, infer mineral compositions, and reconstruct depositional environments with high accuracy (Kim *et al.*, 2020; Fagbemi *et al.*, 2024; Li *et al.*, 2024). Such analyses are essential not only for identifying rock types but also for assessing reservoir quality, structural complexity, and the presence of hydrocarbon-bearing zones (Lai *et al.*, 2024).

To fully harness the potential of geophysical well logs (especially in complex geological environments) advanced analytical techniques are essential. Traditional interpretation methods often fail to capture subtle yet meaningful features embedded in the data. In this context, entropy-based approaches have emerged as powerful tools due to their ability to quantify uncertainty and reveal complexity in time-series datasets. Initially conceptualised in thermodynamics (Clausius, 1865), entropy has since found broad application in domains such as information theory (Mohseni *et al.*, 2024), statistical mechanics (Mageed *et al.*, 2022), biomedical signal analysis (Borowska, 2015), and civil engineering (Amezquita-Sanchez, 2021). Their relevance to analysis of data across a wide range of disciplines lies in their effectiveness at identifying patterns, detecting anomalies, and enhancing interpretability (Amigó *et al.*, 2018; Zhang *et al.*, 2021).

Recent research highlights the adaptability of entropy-based techniques. For example, Huang *et al.* (2024) utilised a combination of entropy measures such as the Shannon entropy and approximate entropy (*ApEn*) to improve the analysis and privacy protection of neurological time-series data. Likewise, entropy metrics have been successfully applied to structural health monitoring, enabling the detection of subtle dynamic changes (Amezquita-Sanchez, 2021). In geoscience, as emphasised by Hristopulos and Varouchakis (2023), maximum entropy methods have become valuable tools for uncovering structure in complex datasets. They are particularly effective for estimating the spectral content of stationary random processes and constructing probabilistic models that account for spatial and temporal dependencies in geological systems.

In geophysics, machine learning (ML) has become an indispensable tool for analysing complex subsurface datasets and uncovering nonlinear patterns that often elude conventional interpretation methods. Applications range from seismic horizon tracking with deep convolutional neural networks (Yang and Sun, 2020) to shear-wave velocity prediction in tight reservoirs using deep learning compared with rock physics models (Fang *et al.*, 2024). Simultaneously, ML has transformed reservoir characterisation by integrating diverse geophysical and petrophysical datasets and detecting complex patterns that often remain inaccessible to traditional statistical approaches (Yu and Ma, 2021; Wang and Chen, 2023; Boulassel *et al.*, 2026). A wide spectrum of ML techniques, spanning supervised algorithms, unsupervised clustering, and deep learning, have proven particularly effective in modelling petrophysical properties and enhancing predictions of lithology, porosity, and permeability. For instance, Singh *et al.* (2020) demonstrated that integrating geophysical well-log data with optimised ML algorithms significantly improves lithology classification accuracy.

Recent studies have further expanded the role of ML in lithology and fluid characterisation. Multifractal and clustering approaches, such as K-means and agglomerative hierarchical clustering, have been successfully applied to reservoir fluid typing (Boulassel *et al.*, 2021), while fractal attribute-based ML workflows have enabled robust fluid prediction in clastic reservoirs (Boulassel *et al.*, 2025a). Ensemble-based regression methods and deep learning models have been developed for predicting or reconstructing dynamic shear wave slowness (Boulassel *et al.*, 2025b, 2025c), and long short-term memory networks have been employed for saturation

prediction from well logs (Cheddad *et al.*, 2025). Additionally, wavelet-based multifractal attributes combined with ML have shown strong potential for oil–gas discrimination (Boulassel *et al.*, 2025d). In lithology studies, feedforward and convolutional neural networks have been widely adopted for facies classification, with promising results (Noroozi and Riahi, 2025).

Moreover, advances in measures of signal complexity and ML frameworks have further enhanced the ability to extract meaningful insights for reservoir evaluation. Entropy and multifractal-based attributes, when integrated with ML classifiers, have been shown to improve fluid typing and lithology discrimination (Ferreira *et al.*, 2009; Hernandez-Martinez *et al.*, 2013; Boulassel *et al.*, 2021; Figiel *et al.*, 2022). These approaches not only strengthen feature extraction but also increase interpretability, thereby providing a more reliable delineation of the lithological boundaries and fluid properties.

Building on this foundation, the present study focuses on optimising sliding parameters for entropy-based analysis of geophysical well logs to enhance lithology discrimination. Using overlapping sliding windows, a common strategy in signal processing, we analyse how variations in window size (*W*) and step size (*S*) affect classification performance. We apply ML classifiers to model lithological patterns based on entropy analysis, aiming to identify the optimal combination of parameters (*W, S*) and algorithms for improved predictive accuracy.

To support the statistical generalisability and robustness of the proposed entropy-based lithology discrimination approach, this study employs a large and diverse dataset. From over 130 wells initially analysed, 117 were selected based on the availability of comprehensive and high-quality data. This resulted in the extraction and utilisation of 1,170 geophysical well logs encompassing a wide range of measurements, including gamma ray (GR), sonic travel time (DTC), neutron porosity (NPHI), RHOB, photoelectric factor (PE), and multiple resistivity logs [RT10 (micro-resistivity), RT20 (very shallow resistivity), RT30 (shallow resistivity), RT60 (medium resistivity), and RT90 (deep resistivity)]. The extensive dataset provides a solid foundation for reliable model training, validation, and interpretation.

This paper is structured as follows: Section 2 outlines the materials and methods, including dataset selection, preprocessing, and modelling approaches. Section 3 presents the results and discusses the impact of sliding parameters on classification outcomes. Ultimately, Section 4 concludes the study, summarising key findings and suggesting avenues for future research.

2. Materials and methods

This study is grounded in data from 117 wells distributed across Algeria's major producing basins, capturing a wide range of geological, lithological, and petrophysical contexts. Tectonically, Algeria is divided into two primary domains by the South Atlas Fault. To the north lies the Alpine domain, shaped by Alpine orogenic processes, while to the south extends the Saharan Platform, a stable geological province underlain by a Precambrian basement and overlain by thick Palaeozoic to Mesozoic sedimentary successions hosting prolific petroleum systems (Zeroug *et al.*, 2007).

From a lithological perspective, Algeria's stratigraphy is characterised by thick Palaeozoic successions dominated by alternating sandstone and shale formations, with additional carbonate and evaporitic intervals. Cambrian to Carboniferous units include extensive sandstone reservoirs interbedded with shale, while the Silurian shale represents a major regional source rock. The Mesozoic succession comprises Triassic red beds and evaporites, Jurassic carbonates, and Cretaceous clastic–carbonate sequences, overlain by Tertiary deposits of mixed lithology. This vertical organisation of reservoir, seal, and source rocks governs the petrophysical responses

observed in the studied well logs (Zeroug *et al.*, 2007).

The hydrocarbon domain is subdivided into three petroleum provinces (eastern, western, and northern) each containing multiple sedimentary basins with distinct structural and stratigraphic settings (Fig. 1). Within this framework, the dataset includes wells from Illizi (13 wells), Berkine and Oued Mya (12 wells each), as well as the Amguid–Messaoud, Hassi Messaoud, Ahnet, Hassi R'Mel, Béchar–Oued Namous, Reggane, Tindouf, and Sbâa basins, which are evenly represented. Although the wells are not confined to a single field-scale work area, all are located within the Saharan Platform petroleum province and share comparable depositional environments and petrophysical characteristics.

Stratigraphically, approximately 55% of the logs correspond to Palaeozoic formations, 35% to Mesozoic units, and 10% to Cenozoic deposits. Lithologically, the dataset is dominated by shale (~61%), followed by sandstone (~20%) and siltstone (~13%), while carbonate lithologies form a minor component (~5%, mainly limestone with minor dolomite). Accordingly, the target formations investigated in this study correspond to heterogeneous siliciclastic reservoirs dominated by shaly sandstones, with shale and siltstone interbeds and limited carbonate contributions. The study, therefore, focuses on lithology discrimination within clastic reservoir systems rather than purely carbonate reservoirs, reflecting typical Palaeozoic–Mesozoic petroleum systems of the Saharan Platform (Zegagh *et al.*, 2025).

Building on this geological and dataset framework, the methodology adopted in this study (Fig. 2) is grounded in a systematic and data-driven approach to optimise entropy-based lithology discrimination models using geophysical well logs. The process begins with the collection

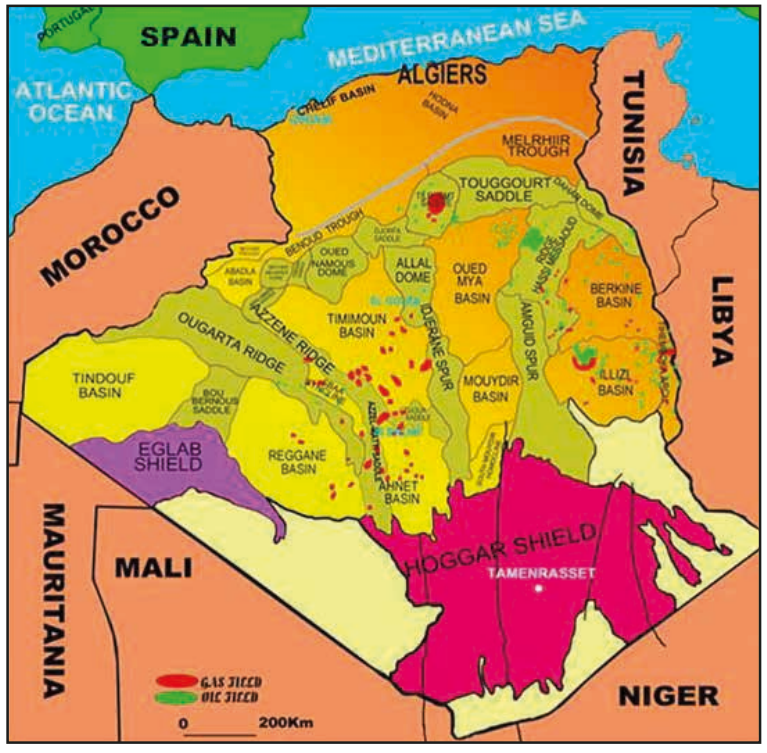


Fig. 1 - Geological map of the sedimentary basins of the Saharan Platform, Algeria (Ministry of Hydrocarbons and Mines; n.d.). The current official ministry site contains the sedimentary-basins/petroleum-provinces information and map.

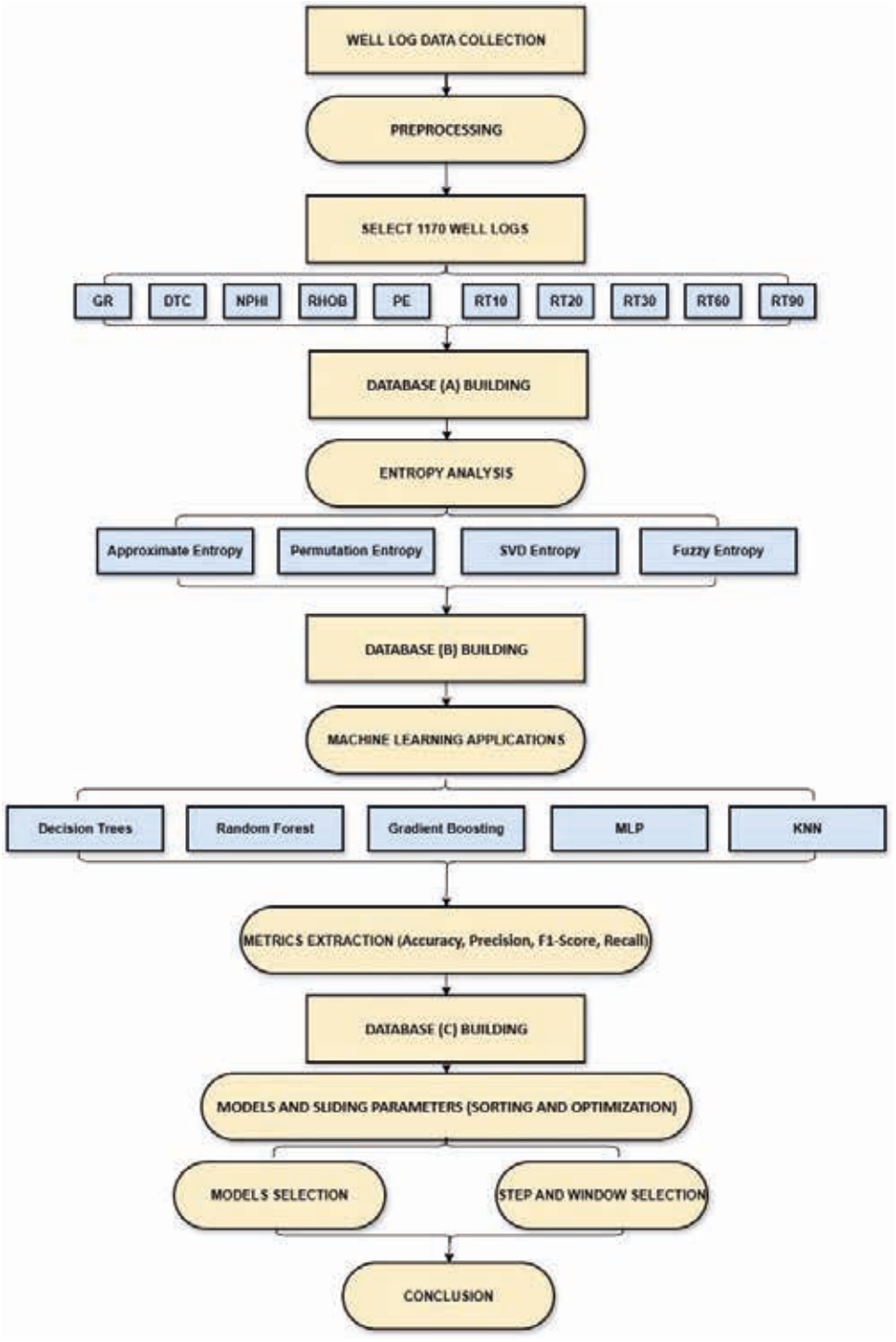


Fig. 2 - The adopted workflow for enhancing entropy-based lithology discrimination through sliding parameters. This figure illustrates the systematic workflow adopted in this study to optimise the sliding parameters, *S* and *W*, for improved lithology discrimination using entropy-based analysis. The methodology outlines a sequential approach, beginning with the collection of geophysical well-log data, followed by comprehensive data preprocessing, entropy computation, and the application of ML algorithms. The workflow is designed to fine-tune the sliding parameters, thereby maximising the effectiveness of entropy measures in distinguishing lithological variations.

and preprocessing of well-log data, where over 130 wells were initially analysed. From these, 117 wells were selected based on the availability of comprehensive and high-quality data, resulting in the extraction and utilisation of 1,170 geophysical well logs in total. The selected logs encompass a wide range of geophysical measurements, including GR, DTC, NPHI, RHOB, PE, and the five resistivity logs (RT10, RT20, RT30, RT60, RT90). These raw logging data underwent rigorous preprocessing (i.e. cleaning, normalisation, and structuring) to form a high-dimensional database (Database_A), which serves as the foundation for all subsequent entropy-based feature extraction and classification analyses.

Entropy-based techniques provide a powerful way to quantify the complexity, irregularity, and information content embedded within geophysical signals. When applied to well-log data, these methods capture subtle variations and hidden patterns that are often difficult to detect using traditional statistical measures, thereby offering valuable insights into lithological heterogeneity. Building on this foundation, the next phase of the workflow involves the application of entropy techniques to the preprocessed well-log data through custom Python scripts and tools such as NeuroKit2, a Python toolbox for signal processing (Makowski *et al.*, 2021). The analysis specifically focuses on exploring the impact of varying sliding parameters (*W* and *S)* on the quality and informativeness of the generated entropy profiles. The entropy-based features extracted during this stage were systematically stored in Database_B, which serves as an organised repository for managing and analysing the generated profiles. Multiple entropy techniques, including *ApEn*, permutation entropy (*PermEn*), singular value decomposition entropy (*SVDEn*), and fuzzy entropy (*FuzzyEn*), were systematically applied to the well logs using a range of sliding parameter combinations. This iterative process allowed for the generation of diverse entropy profiles, enabling a comprehensive evaluation of how different *W*s and *S*s influence the extraction of meaningful geological features. By analysing these profiles, the study aimed to identify the optimal sliding parameters that maximise the discriminative power of the entropy-based features for lithology distinction.

To determine the most appropriate sliding parameters, the study employed a suite of ML models, including DT, RF, GB, MLP, and KNN. These models were trained and evaluated using the entropy-based features derived from the well logs and sliding parameters, with performance metrics such as accuracy, precision, recall, and F1-score used to quantify their effectiveness. The results of these evaluations were systematically compiled into a structured database (Database_C), which facilitated comparative analysis and optimisation. This database enabled the ranking of different sliding parameter combinations based on their performance, ensuring that the selected parameters were not only optimal for the current dataset but also generalisable to other well-log analyses.

The final stage of the methodology involved the optimisation and selection of the best-performing ML models and their corresponding sliding parameters. By systematically sorting and analysing the results in Database_C, the study identified the *W*s and *S*s that yielded the most informative and discriminative entropy profiles. These optimal parameters were validated to ensure their robustness and applicability across diverse geological settings, providing a reliable framework for generating entropy-based features in similar applications.

This methodology combines rigorous data preprocessing, entropy-based feature extraction, and ML optimisation to identify the most appropriate sliding parameters for generating highly informative and generalisable entropy profiles. By leveraging statistical analysis and performance metrics, the study ensures that the resulting entropy-based features enhance lithology discrimination and can be effectively applied to a wide range of geophysical well-log analyses, advancing the accuracy and reliability of subsurface reservoir characterisation.

2.1. Data collection, preprocessing, and selection

The dataset used in this study is derived from geophysical well logs collected from an initial set of over 130 wells, from which 117 wells were selected based on the availability of high-quality and comprehensive data. These wells contain 1,170 geophysical well-logging curves, capturing a diverse range of geophysical and petrophysical properties essential for characterising the subsurface, identifying lithological variations, and estimating hydrocarbon potential.

2.1.1. Data collection

The initial stage of this study involves collecting geophysical well-log measurements from multiple wells across various sedimentary basins. These wells represent diverse geological contexts and petrophysical characteristics. By incorporating data from a wide range of sources and origins, the approach aims to enhance the statistical robustness and generalisability of the results obtained.

2.1.2. Preprocessing

The preprocessing stage ensures that the geophysical well-log data is clean, standardised, and reliable before entropy feature extraction and ML model training. Given the variability in geophysical well-log measurements, preprocessing is essential to eliminate errors, normalise scales, and improve data integrity.

- Feature normalisation. Since well-log values can vary significantly in scale, normalisation techniques such as min-max scaling or z-score standardisation are applied to ensure uniformity across all features.
- Outlier detection and removal. Anomalies in the dataset are identified and filtered using a combination of statistical and ML-based approaches. The preprocessing methods listed are the z-score, Isolation Forest, One-Class Support Vector Machine, Local Outlier Factor, and density-based spatial clustering of applications with noise methods. The results from these methods are aggregated to ensure data reliability and retain only high-quality data for analysis.

2.1.3. Selection of geophysical well-log types

From the 117 wells, a subset of 1,170 geophysical well-logging curves, representing key properties, is chosen. These include:

- GR: measures the natural radioactivity of rock formations, helping distinguish between different lithologies, such as shale and sandstone;
- DTC: provides acoustic velocity data, which is useful for estimating porosity and identifying lithology characteristics;
- NPHI: estimates rock porosity by analysing neutron interactions with hydrogen atoms, often used alongside density logs for more accurate assessments;
- RHOB: measures rock density, a key factor in determining porosity and classifying lithology.
- PE: helps assess lithology composition by analysing how rocks absorb X-rays;
- resistivity logs: measure the electrical resistivity of rock formations at different depths, crucial for distinguishing between hydrocarbon-bearing and water-filled zones.

2.2. Structured database (Database_A)

The geophysical well-log data is structured in Database_A (Table 1) to improve accessibility, maintain consistency, and support future expansion, forming a strong basis for entropy-based feature extraction and ML modelling.

Table 1 - Sample of preprocessed geophysical well logs from Database_A. This table presents selected geophysical well logs from Database_A after preprocessing, including cleaning, normalisation, and formatting. Featured logs include GR, DTC, NPHI, RHOB, PE, and resistivity logs (RT10, RT20, RT30, RT60, RT90).

WellName	DEPTH (m)	DTC (µs/ft)	GR (gAPI)	NPHI (v/v)	PE (b/e)	RHOB (g/cc)	RT10 (ohm.m)	RT20 (ohm.m)	RT30 (ohm.m)	RT60 (ohm.m)	RT90 (ohm.m)
WELL-1	2022	97.25523088	238.3980966	0.4050964	4.804556	2.400956	14.32780855	17.723535	19.88951841	16.46595101	19.51017798
WELL-1	2022.1524	96.96248491	242.3992532	0.4142139	4.80458	2.402756	14.01547554	17.99795251	20.7349686	23.21117467	18.74643256
WELL-1	2022.3048	97.3301261	242.2910228	0.4135328	4.795811	2.405067	14.72543524	19.14053701	22.42691819	30.27978673	20.59294411
WELL-1	2022.4572	98.6727305	244.7717872	0.3990474	4.769174	2.408901	16.54920626	21.43466443	25.37465951	32.55003076	23.29216075
WELL-1	2022.6096	100.0160974	242.895992	0.3936484	4.713513	2.413908	17.30125051	21.28645844	25.36256825	30.51935952	21.92040466
WELL-1	2022.762	100.6003236	236.8751417	0.3837448	4.62815	2.388613	17.93633753	20.7288116	24.15418318	26.97246962	20.38360537
WELL-1	2022.9144	101.0153185	232.0190201	0.3876831	4.524084	2.336008	19.74232319	20.68175269	24.50684678	24.53274252	20.73132901
WELL-1	2023.0668	101.2304208	234.554268	0.3851008	4.412728	2.307636	22.66792342	21.09853655	26.08993466	25.73867748	22.08018927
WELL-1	2023.2192	101.0107547	240.4092617	0.3828858	4.32409	2.359081	20.84168854	22.79080473	28.28687061	25.61605582	24.61544134
WELL-1	2023.3716	100.8733988	241.871396	0.3747746	4.289556	2.477622	20.42372118	23.565848	26.9130942	22.34026873	24.39679916
WELL-1	2023.524	100.2964461	241.7621192	0.3753647	4.335214	2.574632	20.31585891	22.81620356	24.91985053	22.23354247	22.00450295
WELL-1	2023.6764	99.39985598	244.2319537	0.3665735	4.400935	2.620042	19.60898388	21.10611109	24.81909645	23.14966453	21.95651211
WELL-1	2023.8288	98.5805474	240.681101	0.3683898	4.42349	2.627565	20.07405857	20.31782066	24.14583022	23.25337653	21.1729847
WELL-1	2023.9812	95.55772723	231.3991045	0.3665141	4.388764	2.632555	20.3760779	19.64757734	21.59999186	19.33294134	17.12487068
WELL-1	2024.1336	95.41056859	216.9101198	0.3747381	4.322052	2.645726	17.58936726	17.03270337	23.88529272	21.41339251	19.09664153
WELL-1	2024.286	94.91924538	210.1553211	0.3807207	4.295258	2.649564	19.56778444	17.74901147	22.67776941	22.72074238	20.3421573
WELL-1	2024.4384	94.76533844	210.3288298	0.372612	4.308948	2.642653	22.02265494	17.80928858	20.99267561	19.1482465	19.49037122
WELL-1	2024.5908	95.02423541	213.6225034	0.362538	4.346252	2.632369	23.5963411	17.33940249	20.87371917	20.99973389	18.58001832
WELL-1	2024.7432	95.43478823	212.3608625	0.3550133	4.376525	2.627609	24.16696884	16.87978155	22.34731776	23.31731725	21.2271212
WELL-1	2024.8956	95.70709268	205.9909955	0.3675196	4.393899	2.626153	25.74560169	16.91704137	22.90158459	20.67643711	22.1863227
WELL-1	2025.048	96.12327142	206.5909505	0.3788504	4.412395	2.619717	27.85478785	17.3344593	22.07928843	21.00253687	20.85327485
WELL-1	2025.2004	96.42463957	214.4902032	0.3836854	4.423498	2.610734	31.96816502	17.504578	22.73017769	23.54684438	21.80776594
WELL-1	2025.3528	96.40232931	226.8589631	0.3905525	4.408276	2.603425	30.22724358	17.9531446	22.71467833	23.45927128	21.47099089
WELL-1	2025.5052	96.33175844	227.6669963	0.3841398	4.363914	2.600507	29.49274128	18.33512436	21.55767427	22.5817158	19.89031059
WELL-1	2025.6576	96.3007807	223.8442456	0.3754664	4.288719	2.604611	29.4635727	18.16402538	22.10216079	22.51390595	20.85759756
WELL-1	2025.81	96.18551457	221.6382979	0.3602374	4.238622	2.611953	29.16201263	17.16077478	22.44692092	22.95957812	21.19286378
WELL-1	2025.9624	95.58491259	218.5700793	0.3624963	4.214679	2.624368	27.84127466	15.77689634	21.56055201	21.00307154	20.98155752
WELL-1	2026.1148	95.34539181	220.7629939	0.3696202	4.206451	2.637245	26.46925726	14.67442923	21.70872765	20.5110721	20.76603395

2.3. Application of entropy techniques to well-log data

2.3.1. Sliding window segmentation

Following preprocessing and data selection, each geophysical well log is partitioned into overlapping depth intervals using a sliding-window scheme. Two parameters govern this segmentation: *W*, which sets how much contextual information from the signal is captured within each interval, and *S*, which controls the advance between consecutive windows and, therefore, the degree of overlap and the vertical resolution of the resulting attribute profiles (Fig. 3).

Because these parameters directly shape the density, stability, and interpretability of the entropy-based attributes, we deliberately explored a wide range of *W and S* configurations and systematically evaluated them using performance indicators from ML models for lithological characterisation and distinction based on the entropy attributes. The rationale was twofold: i) to identify parameter combinations that yield the best predictive performance and ii) to ensure robustness by covering all practical scenarios, i.e. high *W*–high *S*, high *W*–low *S*, low *W*–high *S*, and low *W*–low *S*. This exhaustive screening represents the most rigorous methodological path to minimise bias from arbitrary choices and to generalise across diverse geological conditions.

The final set of parameter combinations was, therefore, guided by both theory and evidence.

From a theoretical perspective, larger windows capture broader depositional context but smooth fine features, while smaller windows accentuate thin beds and sharp transitions. Likewise, smaller steps produce denser, higher-resolution profiles at the cost of greater sensitivity to noise, whereas larger steps yield smoother, more conservative profiles. To avoid bias and to reflect best practices in signal investigation, time-series analysis and geophysical well-log processing, we examined a broad and balanced range of values so that both fine-scale (small *W*, small *S*) and coarse-scale (large *W*, large *S*) characterisations were represented.

Operationally, windows were required to be fully contained within valid data; entropy attributes were computed only when a minimum proportion of in-window samples was available to avoid artefacts from gaps. Each window's attribute value was assigned to the window centre depth to produce continuous profiles. To make the impact of the sliding parameters explicit, Table 2 reports representative (*W, S, IRI*$_{\%}$) combinations, and Fig. 2 illustrates how these choices control the number and displacement of overlapping segments along the depth axis. In total, about 136 configurations were applied to 1,170 well logs to compute four entropy profiles per log, providing a comprehensive basis for comparison. Because larger *W* and *S* reduce the number of extracted samples, a careful selection of sliding parameters is essential to preserve informativeness and ensure meaningful geological interpretation. The resulting sampling density is summarised by the informational richness indicator (*IRI)* or its normalised form (*IRI*$_{\%}$), which is developed further in the next subsection.

2.3.2. Informational richness indicator

To quantitatively assess the amount of information extracted from a signal using a sliding window approach, we introduce the *IRI*. This metric captures the number of segments generated when systematically partitioning the signal, thereby reflecting the level of detail and granularity embedded in the transformed dataset.

When a sliding window of size *W* is applied to a signal of length *N*, moving forwards by a step *S*, the number of resulting segments (or samples) can be calculated as:

$$IRI\ (N,W,S)\ =\ N_{Extracted_samples}\ =\ \frac{N-W}{S}+1\ . \tag{1}$$

Eq. (1) provides the absolute count of extracted segments. However, because signals may differ in length, it is useful to express this measure in relative terms. Normalising by the total number of original samples yields the percentage-based form of *IRI*:

$$IRI_{\%}\ =\ \frac{N_{Extracted_samples}}{N}\ \ *\ \ 100\ \ \%\ . \tag{2}$$

By substituting Eq. (1) in Eq. (2), a more compact formulation emerges:

$$IRI_{\%}\ =\ \frac{N+S-W}{N*S}\ \ *\ \ 100\ \ \% \tag{3}$$

where *N* is the total number of original signal samples, *W* is the window size, *S* is the step size, and *IRI* equals the total number of extracted samples ($N_{Extracted_samples}$). $IRI_{\%}$ is the proportion of

extracted samples relative to the original signal, expressed as a percentage.

This formulation explicitly captures the relationship between the original signal and the sliding window parameters. The logic is straightforward:

1. windowing the signal. A window of size *W* scans the signal, starting at the first sample, shifting forwards by *S* samples at each time;
2. counting the windows. The total number of valid positions for the window is determined by how many full windows can fit into the signal length *N*, considering the *S*;
3. quantifying informational richness. The resulting *IRI* reflects the number of segments, and, thus, the effective resolution, produced under the chosen configuration.

Table 2 - Example of sliding window configurations used in this study for well logs with N = 519. The table shows combinations of W, S, and the resulting $IRI_{\%}$, which defines the degree of overlapping segmentation of the dataset.

Combination (S, W, IRI, IRI%)	Sliding Window (W)																
	W=10	W=20	W=30	W=40	W=50	W=60	W=70	W=80	W=90	W=100	W=110	W=120	W=130	W=140	W=150	W=200	W=300
Sliding Step (S) S=1	(1, 10,98.27)	(1, 20,96.34)	(1, 30,94.41)	(1, 40,92.48)	(1, 50, 90.56)	(1, 60, 88.63)	(1, 70, 86.70)	(1, 80, 84.78)	(1, 90, 82.85)	(1, 100, 80.92)	(1, 110, 78.99)	(1, 120, 77.07)	(1, 130, 75.14)	(1, 140, 73.21)	(1, 150, 71.28)	(1, 200, 61.66)	(1, 300, 42.39)
S=2	(2, 10, 49.13)	(2, 20, 48.17)	(2, 30, 47.21)	(2, 40, 46.26)	(2, 50, 45.30)	(2, 60, 44.33)	(2, 70, 43.36)	(2, 80, 42.39)	(2, 90, 41.43)	(2, 100, 40.46)	(2, 110, 39.50)	(2, 120, 38.53)	(2, 130, 37.57)	(2, 140, 36.61)	(2, 150, 35.65)	(2, 200, 30.83)	(2, 300, 21.19)
S=3	(3, 10, 32.75)	(3, 20, 32.18)	(3, 30, 31.60)	(3, 40, 31.02)	(3, 50, 30.44)	(3, 60, 29.87)	(3, 70, 29.29)	(3, 80, 28.71)	(3, 90, 28.13)	(3, 100, 27.55)	(3, 110, 26.97)	(3, 120, 26.40)	(3, 130, 25.82)	(3, 140, 25.24)	(3, 150, 24.66)	(3, 200, 20.62)	(3, 300, 14.07)
S=4	(4, 10, 24.66)	(4, 20, 24.28)	(4, 30, 23.89)	(4, 40, 23.51)	(4, 50, 22.93)	(4, 60, 22.55)	(4, 70, 22.16)	(4, 80, 21.78)	(4, 90, 21.40)	(4, 100, 21.00)	(4, 110, 20.62)	(4, 120, 20.23)	(4, 130, 19.85)	(4, 140, 19.46)	(4, 150, 19.08)	(4, 200, 15.61)	(4, 300, 10.79)
S=5	(5, 10, 19.65)	(5, 20, 19.46)	(5, 30, 19.08)	(5, 40, 18.88)	(5, 50, 18.69)	(5, 60, 18.50)	(5, 70, 18.31)	(5, 80, 18.11)	(5, 90, 17.92)	(5, 100, 17.73)	(5, 110, 17.53)	(5, 120, 17.34)	(5, 130, 17.15)	(5, 140, 16.96)	(5, 150, 16.76)	(5, 200, 13.69)	(5, 300, 9.41)
S=10	(10, 10, 9.82)	(10, 20, 9.72)	(10, 30, 9.52)	(10, 40, 9.43)	(10, 50, 9.33)	(10, 60, 9.24)	(10, 70, 9.15)	(10, 80, 9.05)	(10, 90, 8.96)	(10, 100, 8.86)	(10, 110, 8.77)	(10, 120, 8.67)	(10, 130, 8.57)	(10, 140, 8.48)	(10, 150, 8.38)	(10, 200, 6.84)	(10, 300, 4.70)
S=15	(15, 10, 6.55)	(15, 20, 6.49)	(15, 30, 6.36)	(15, 40, 6.30)	(15, 50, 6.24)	(15, 60, 6.18)	(15, 70, 6.12)	(15, 80, 6.05)	(15, 90, 5.99)	(15, 100, 5.93)	(15, 110, 5.87)	(15, 120, 5.80)	(15, 130, 5.74)	(15, 140, 5.68)	(15, 150, 5.62)	(15, 200, 4.56)	(15, 300, 3.13)
S=20	(20, 10, 4.91)	(20, 20, 4.86)	(20, 30, 4.77)	(20, 40, 4.72)	(20, 50, 4.67)	(20, 60, 4.62)	(20, 70, 4.57)	(20, 80, 4.52)	(20, 90, 4.47)	(20, 100, 4.43)	(20, 110, 4.38)	(20, 120, 4.33)	(20, 130, 4.28)	(20, 140, 4.23)	(20, 150, 4.18)	(20, 200, 3.42)	(20, 300, 2.35)

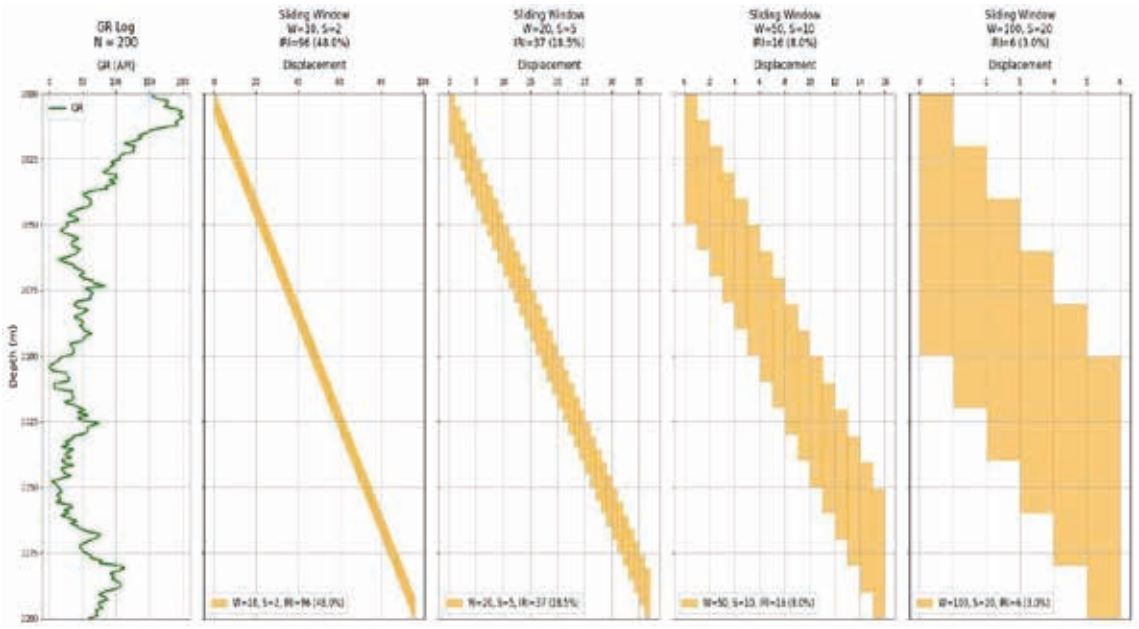


Fig. 3 - Effect of sliding window parameters on geophysical well-log segmentation and *IRI*. This figure illustrates the impact of varying sliding window configurations on the segmentation of GR well log (N = 200 samples). The GR log is displayed on the left, followed by four panels showing displacement patterns resulting from different W and S combinations. Specifically, the configurations include W = 10, S = 2 (*IRI* = 96; $IRI_{\%}$ = 48.0%), W = 20, S = 5 (*IRI* = 37; $IRI_{\%}$ = 18.5%), W = 50, S = 10 (*IRI* = 16; $IRI_{\%}$ = 8.0%), and W = 100, S = 20 (*IRI* = 6; $IRI_{\%}$ = 3.0%). As the W increases and the S becomes coarser, the number of segments decreases and the overlap between windows reduces. This visualisation highlights how parameter choices influence data partitioning, which is critical for entropy-based analysis and ML-driven lithology classification.

A higher *IRI* value indicates that more segments are extracted, leading to finer-scale informational granularity. Conversely, a lower *IRI* reflects coarser sampling, where fewer segments summarise the signal. Thus, the *IRI* offers a simple yet rigorous means of assessing the richness of information derived from sliding window transformations.

In our workflow, the *IRI* plays two complementary roles:

- design metric: guiding the construction of parameter grids that span from fine- to coarse-scale signal characterisations;
- diagnostic metric: contextualising ML performance indicators in lithological characterisation tasks.

By explicitly defining the *IRI* and its properties, we ensure that parameter choices are transparent, reproducible, and comparable across different wells and studies. Configurations achieving a balance between a sufficiently high *IRI* (to capture target features) and strong predictive performance are prioritised, thereby enhancing both methodological robustness and interpretability.

2.3.3. Entropy-based feature extraction

After segmentation, the next essential step is extracting entropy features from each sliding window to enhance the dataset's ability to capture geological variations. By applying entropy measures, we quantify the complexity, randomness, and predictability of geophysical well-log signals, ensuring a more detailed representation of subsurface properties. The entropy measures utilised in this study include:

A. *ApEn: ApEn* is a statistical measure that quantifies the unpredictability or regularity of time-series data. It was introduced by Pincus (1991) initially to analyse chaotic dynamics and later applied to assess complexity in physiological, financial, and other types of data. The fundamental idea behind *ApEn* is to evaluate the probability that patterns in a given time series remain similar at the next incremental data point. Lower *ApEn* values indicate more regular and predictable data, while higher values suggest greater irregularity. *ApEn* is defined as:

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r) \tag{4}$$

where $\phi^m(r)$ quantifies the probability of similar runs of length remaining similar at $m + 1$, m is the embedding dimension, r is the similarity threshold, and N is the length of the time series;

B. *PermEn: PermEn* was introduced by Bandt and Pompe (2002) as a method to quantify the randomness of a time series by analysing the frequency of ordinal patterns. It is particularly useful for detecting dynamical changes in nonlinear and chaotic systems and is given by:

$$PermEn(m) = -\sum_{i=1}^{n!} p_i \log(p_i) \tag{5}$$

where p_i is the probability of each permutation pattern in the time series;

C. *SVDEn: SVDEn* is a quantitative measure that captures the complexity of a dataset by analysing the distribution of its singular values (Roberts *et al.*, 1999; Strydom *et al.*, 2021). By decomposing

a data matrix into its singular components, *SVDEn* assesses how information is distributed across different dimensions, with higher entropy indicating a more uniform and complex structure. It is defined as:

$$SVDEn = -\sum_{i=1}^{k} \sigma_i^2 \, log(\sigma_i^2) \tag{6}$$

where σ_i^2 are the normalised singular values;

D. *FuzzyEn: FuzzyEn*, introduced by De Luca and Termini (1972), quantifies uncertainty in systems with imprecise or noisy data using fuzzy logic principles. By employing gradual membership functions, it captures subtle variations that traditional entropy measures might overlook. *FuzzyEn* is widely applied in physiological signal analysis for its robustness and sensitivity. Mathematically, it is defined as:

$$FuzzyEn(m, r, N) = ln\phi^m(r) - ln\phi^{m+1}(r) \tag{7}$$

Table 3 - Example of entropy-based features extracted from Database_B. This table showcases a sample of entropy-derived attributes calculated from geophysical well logs in Database_B, including *PerEn*, *ApEn*, *FuzzyEn*, and *SVDEn*. Each entry is accompanied by the corresponding well identifier, log type, lithology label, and the sliding window parameters (*W*, *S*, *IRI*, and $IRI_{\%}$) used during feature extraction. These features form a key part of the input for subsequent lithology classification and pattern analysis.

Depth	well	log	PermEntropy	SVDEntropy	ApproximateEntropy	FuzzyEntropy	window	step	IRI	IRI %	Lithology
Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre
2218.5052	WELL-1	DTC	0.676992412660448	0.148398728115805	0.133531392624522	0.782906017723111	10	1	10169	99.91	Shale
2218.6576	WELL-1	DTC	0.676992412660448	0.141616750415546	0.133531392624522	0.787001946685769	10	1	10169	99.91	Shale
2218.81	WELL-1	DTC	0.676992412660448	0.0968745131877661	0.133531392624522	0.757104086525667	10	1	10169	99.91	Shale
2218.9624	WELL-1	DTC	0.543775417182036	0.0554375060562324	0.133531392624522	0.68373577530598	10	1	10169	99.91	Shale
2219.1148	WELL-1	DTC	0.700698027129531	0.0490705698582772	0.133531392624522	0.870641224270585	10	1	10169	99.91	Shale
2219.2672	WELL-1	DTC	0.700698027129531	0.0411270409175828	0.133531392624522	1.24537943312953	10	1	10169	99.91	Shale
2219.4196	WELL-1	DTC	0.737201820799307	0.0310101426411352	0.133531392624522	1.50795475242572	10	1	10169	99.91	Shale
2219.572	WELL-1	DTC	0.833915022607942	0.0239399671060431	0.133531392624522	1.76258490035892	10	1	10169	99.91	Shale
2219.7244	WELL-1	DTC	0.833915022607942	0.0234068180947007	0.133531392624522	1.90299244143137	10	1	10169	99.91	Shale
2219.8768	WELL-1	DTC	0.833915022607943	0.0210799890358146	0.133531392624522	1.41587010454402	10	1	10169	99.91	Shale
2220.0292	WELL-1	DTC	0.773705614469083	0.0161572599577557	0.133531392624522	1.16857312287377	10	1	10169	99.91	Shale
2220.1816	WELL-1	DTC	0.59915567063867	0.0129179321617332	0.133531392624522	0.942155207275477	10	1	10169	99.91	Shale
2220.334	WELL-1	DTC	0.410558421703625	0.0121922060948437	0.133531392624522	0.835930596049479	10	1	10169	99.91	Shale
2220.4864	WELL-1	DTC	0.410558421703625	0.0111824173223519	0.133531392624522	0.813746915791919	10	1	10169	99.91	Shale
2220.6388	WELL-1	DTC	0.410558421703625	0.0122103732751299	0.133531392624522	0.953573879069523	10	1	10169	99.91	Shale
2220.7912	WELL-1	DTC	0.410558421703625	0.027066999524087	0.133531392624522	0.840253419893881	10	1	10169	99.91	Shale
2220.9436	WELL-1	DTC	0.502442468830035	0.0493124951749345	0.0397554025154641	0.571434448788581	10	1	10169	99.91	Shale
2221.096	WELL-1	DTC	0.543775417182036	0.0643551536695408	0.0601962661562963	0.606324335547913	10	1	10169	99.91	Shale
2221.2484	WELL-1	DTC	0.700698027129531	0.0663559436520507	0.0405781895729376	0.748778277160739	10	1	10169	99.91	Shale
2221.4008	WELL-1	DTC	0.700698027129531	0.0722141737798522	0.0209601129895782	1.17401515825459	10	1	10169	99.91	Shale
2221.5532	WELL-1	DTC	0.700698027129531	0.0816872563827588	0.0397554025154641	1.52912172965779	10	1	10169	99.91	Shale
2221.7056	WELL-1	DTC	0.700698027129531	0.0882324293472628	0.133531392624522	1.38935935415395	10	1	10169	99.91	Shale
2221.858	WELL-1	DTC	0.700698027129531	0.0904163584893762	0.133531392624522	1.36301590110729	10	1	10169	99.91	Shale
2222.0104	WELL-1	DTC	0.700698027129531	0.0873818224647979	0.133531392624522	1.26880310944874	10	1	10169	99.91	Shale

2.3.4. Entropy-based feature database (Database_B)

As a critical step in the workflow, Database_B (Table 3) was constructed to store the entropy-based attributes derived from the segmented geophysical well-log data (Fig. 4). Each entry in this database includes multiple entropy measures, i.e. *ApEn*, *PermEn*, *SVDEn*, and *FuzzyEn*, calculated over sliding windows defined by specific *W* and *S* parameters. In addition to the entropy values, Database_B integrates essential contextual information such as well identifiers, log types, depth, *IRI*, $IRI_{\%}$, and lithology. This curated dataset serves as a robust foundation for training and evaluating ML models.

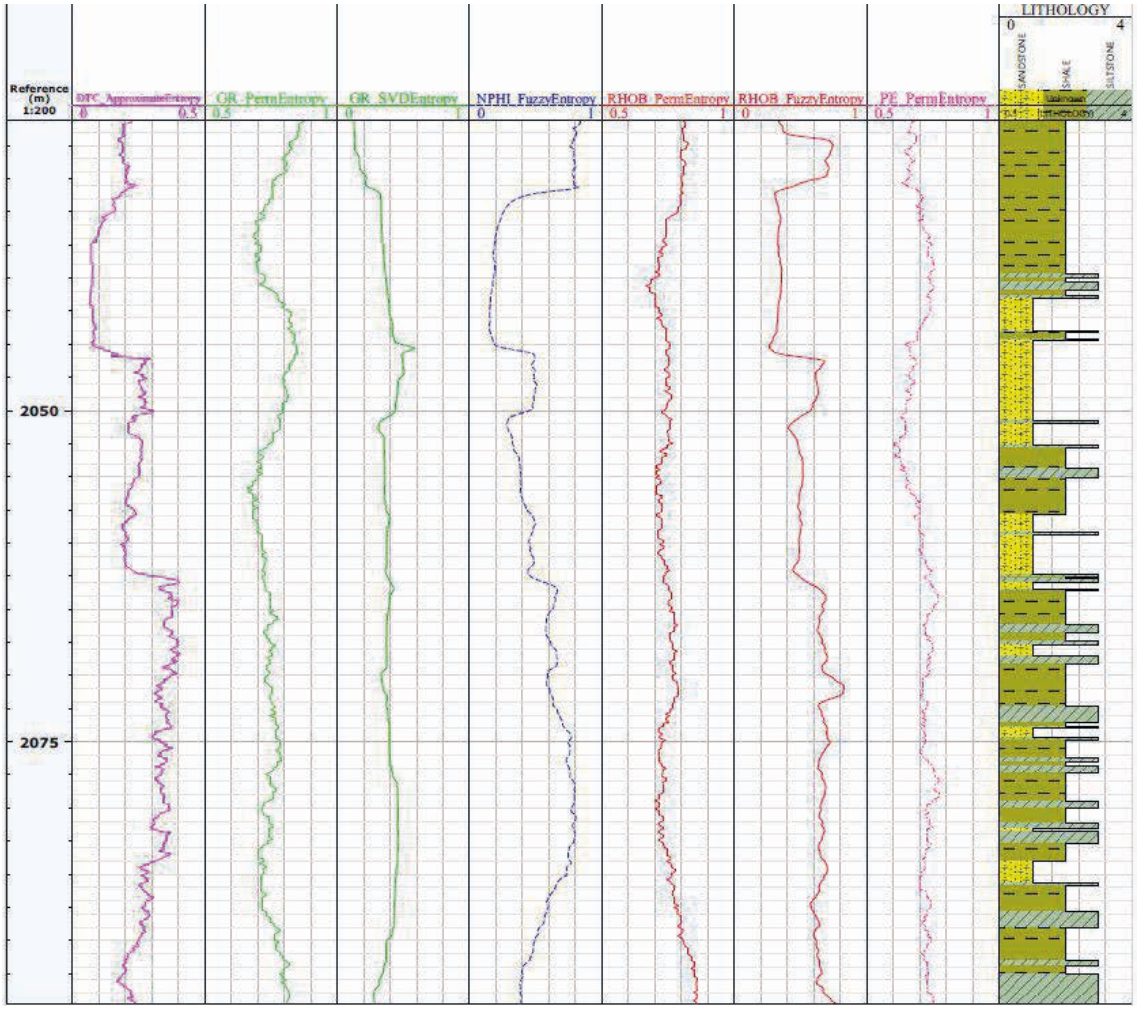


Fig. 4 - Example of entropy-based complexity profiles of some geophysical well logs. This figure presents entropy-based complexity profiles derived from geophysical well-log data, utilising four distinct entropy measures: *ApEn*, *FuzzyEn*, *PermEn*, and *SVDEn*. The analysis spans multiple geophysical logs, including GR, DTC, NPHI, and RHOB. The resulting profiles highlight subtle variations in signal complexity with depth, providing valuable insights into subsurface heterogeneity. The lithological classification is shown on the far right for correlation with the complexity profiles.

2.3.5. ML model applications

Entropy profiles were used as input features (*X*), while lithology classes constituted the target variable (*Y*). Five widely used supervised ML models were employed to evaluate lithology discrimination performance: KNN, RF, GB, MLP, and DT. Only a concise description of each method is provided here, as these algorithms are well established in the literature.

KNN classifies samples based on the majority class of their *k* nearest neighbours in feature space, enabling flexible modelling of nonlinear patterns without explicit training (Cover and Hart, 1967). RF is an ensemble method that aggregates predictions from multiple DTs trained on randomly selected subsets of data and features, improving robustness and reducing overfitting (Breiman, 2001). GB constructs predictive models sequentially by minimising a loss function through additive weak learners, typically DTs (Friedman, 2001). The MLP is a feedforward neural network composed of multiple hidden layers that model complex feature interactions using nonlinear activation functions (Rosenblatt, 1958; Goodfellow *et al.*, 2016). DTs partition the feature space using recursive splits that maximise information gain, providing interpretable classification rules (Quinlan, 1986).

Model performance was evaluated using standard classification metrics derived from the confusion matrix, including accuracy, precision, recall, and the F1-score. The confusion matrix (Fig. 5) summarises correct and incorrect predictions across classes and forms the basis for all reported metrics (Sokolova *et al.*, 2009; Powers, 2020). Accuracy measures the proportion of correct predictions, precision quantifies the reliability of positive predictions, recall evaluates the ability to detect true lithologies, and the F1-score provides a balanced measure that accounts for both false positives and false negatives. These metrics collectively offer a comprehensive assessment of classification effectiveness, particularly in the presence of class imbalance.

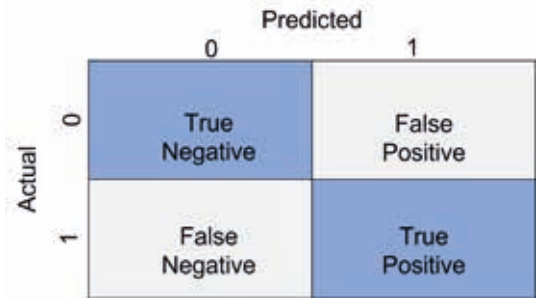


Fig 5 - Standard layout of a 2×2 confusion matrix.

2.3.6. Performance storage and comparative analysis (Database_C)

Following the training and evaluation of ML models, all classification performance metrics were systematically stored in Database_C (Table 4). This database served as a structured repository for key evaluation parameters, including:

- classifier type (e.g. RF, MLP, GB, KNN, DT);
- sliding *W*;
- sliding *S*;
- accuracy, precision, recall, and F1-score.

By organising the results in Database_C, we enabled a direct comparison of model effectiveness across different configurations, providing a robust foundation for selecting optimal parameters.

Table 4 - Example of classification performance metrics on Database_C, this table showcases the performance metrics of various ML models evaluated on Database_C, across a range of sliding window parameters (*W*, *S, IRI*, and $IRI_{\%}$). The reported metrics (accuracy, precision, recall, and F1-score) were systematically assessed to facilitate optimal model selection and parameter tuning. This comparative analysis provides valuable insights into how different configurations impact classification effectiveness.

Classifier	Accuracy	Precision	Recall	F1-Score	Window	Step	IRI	IRI %
Random Forest	0.8964	0.8929	0.8964	0.8943	80	1	10099	99.22
Random Forest	0.8957	0.8908	0.8957	0.8902	300	1	9879	97.06
Random Forest	0.8929	0.8868	0.8929	0.8880	100	1	10079	99.03
Random Forest	0.8926	0.8851	0.8926	0.8854	60	1	10119	99.42
Random Forest	0.8923	0.8871	0.8923	0.8887	150	1	10029	98.54
Random Forest	0.8914	0.8856	0.8914	0.8873	140	1	10039	98.63
Random Forest	0.7747	0.7511	0.7747	0.7590	150	3	3344	32.86
Random Forest	0.8903	0.8835	0.8903	0.8843	30	1	10149	99.72
MLP	0.7323	0.7032	0.7323	0.7140	300	15	660	6.48
Random Forest	0.8889	0.8828	0.8889	0.8835	130	1	10049	98.73
Random Forest	0.8883	0.8812	0.8883	0.8828	90	1	10089	99.13
Gradient Boosting	0.7204	0.6714	0.7204	0.6735	70	10	1012	9.94
Random Forest	0.8841	0.8771	0.8841	0.8786	200	1	9979	98.04
KNN	0.6579	0.5936	0.6579	0.6167	80	20	506	4.97
Random Forest	0.7854	0.7605	0.7854	0.7701	300	4	2471	24.28
KNN	0.8736	0.8690	0.8736	0.8706	80	1	10099	99.22
Random Forest	0.8707	0.8619	0.8707	0.8601	20	1	10159	99.81
KNN	0.8121	0.7980	0.8121	0.8019	70	2	5055	49.67
Random Forest	0.7868	0.7602	0.7868	0.7649	110	5	2015	19.80
KNN	0.8640	0.8558	0.8640	0.8570	300	1	9879	97.06
KNN	0.8638	0.8564	0.8638	0.8573	120	1	10059	98.83
KNN	0.8628	0.8562	0.8628	0.8587	70	1	10109	99.32
KNN	0.8611	0.8518	0.8611	0.8533	100	1	10079	99.03
Decision Tree	0.8607	0.8594	0.8607	0.8594	300	1	9879	97.06
KNN	0.8606	0.8534	0.8606	0.8555	140	1	10039	98.63
KNN	0.8597	0.8516	0.8597	0.8526	200	1	9979	98.04
KNN	0.8577	0.8485	0.8577	0.8496	110	1	10069	98.93
Decision Tree	0.8571	0.8572	0.8571	0.8570	150	1	10029	98.54

3. Results

This section presents the quantitative outcomes obtained from the classification experiments conducted under varying *W*, *S*, and $IRI_{\%}$ configurations

3.1. Effect of window size on classification performance

3.1.1. Accuracy

Classification accuracy increases with *W* for all classifiers (Fig. 6). Small windows (*W* = 10) produce lower accuracy values, ranging from approximately 0.65 for DT to 0.73 for RF. Accuracy improves markedly between *W* = 10 and *W* = 60, reaching values above 0.80 for the best-performing classifiers. Beyond *W* ≈ 100, accuracy gains become marginal and tend to stabilise.

RF consistently achieves the highest accuracy across all *W*s, with peak accuracy exceeding 0.85, followed by GB and MLP. KNN shows moderate improvement, while DT exhibits the lowest accuracy and highest variability.

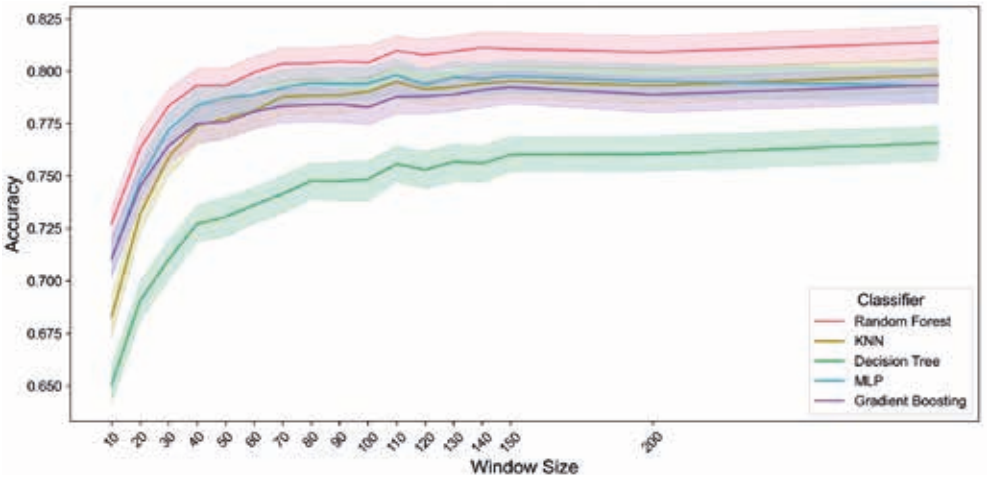


Fig. 6 - Effect of *W* on classification accuracy. This figure illustrates the variation in classification accuracy for different ML models as a function of sliding *W*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

3.1.2. Precision

Precision improves steadily as *W* increases (Fig. 7). At small *W*s (*W* ≤ 30), all classifiers show relatively low precision, typically below 0.75. A pronounced increase occurs between *W* = 30 and *W* = 60, after which improvements diminish. For *W*s above 100–150, precision curves plateau,

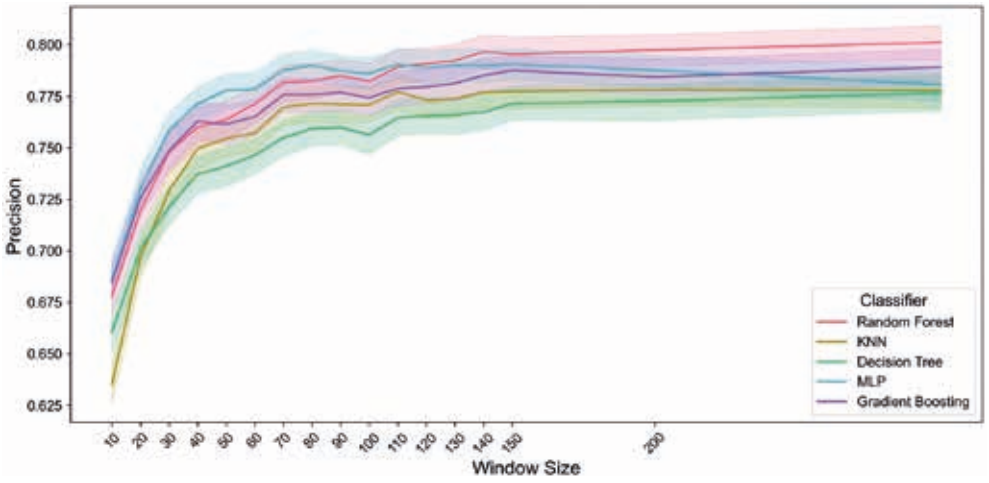


Fig. 7 - Effect of *W* on classification precision. This figure illustrates the variation in classification precision for different ML models as a function of sliding *W*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

with maximum values approaching 0.85. RF consistently attains the highest precision, followed closely by MLP and GB. KNN and DT remain less stable, with DT showing the largest variability.

3.1.3. F1-score

The F1-score exhibits a clear upward trend with increasing *W* (Fig. 8). The most significant improvement occurs between *W* = 10 and *W* = 60, where the F1-score increases from values near 0.70 to above 0.82 for the strongest classifiers. Beyond *W* = 100, F1-score values stabilise, indicating diminishing returns. RF achieves the highest and most consistent F1-scores, with peak values exceeding 0.85, while GB and MLP remain competitive. KNN and DT perform less favourably, with DT displaying substantial variability.

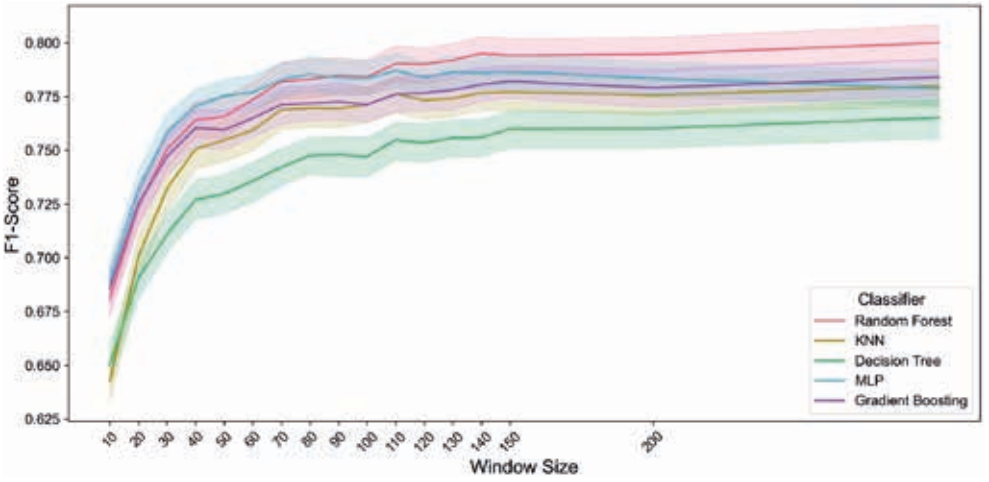


Fig. 8 - Effect of *W* on classification F1-score. This figure illustrates the variation in classification F1-score for different ML models as a function of sliding *W*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

3.1.4. Recall

Recall increases with *W* for all classifiers (Fig. 9), particularly between *W* = 10 and *W* = 60, where values rise from approximately 0.68 to above 0.80. After *W* ≈ 100, recall curves flatten. RF maintains the strongest recall throughout, with values exceeding 0.85, followed by MLP and KNN. GB and DT show lower recall and greater variability.

3.2. Effect of step size on classification performance

3.2.1. Accuracy

Accuracy decreases as *S* increases (Fig. 10). The sharpest decline occurs between *S* = 1 and *S* = 3, where accuracy drops by approximately 5–8% across classifiers. RF consistently outperforms other classifiers, maintaining accuracy above 0.80 at small *S*s, while DT exhibits the most pronounced degradation and variability.

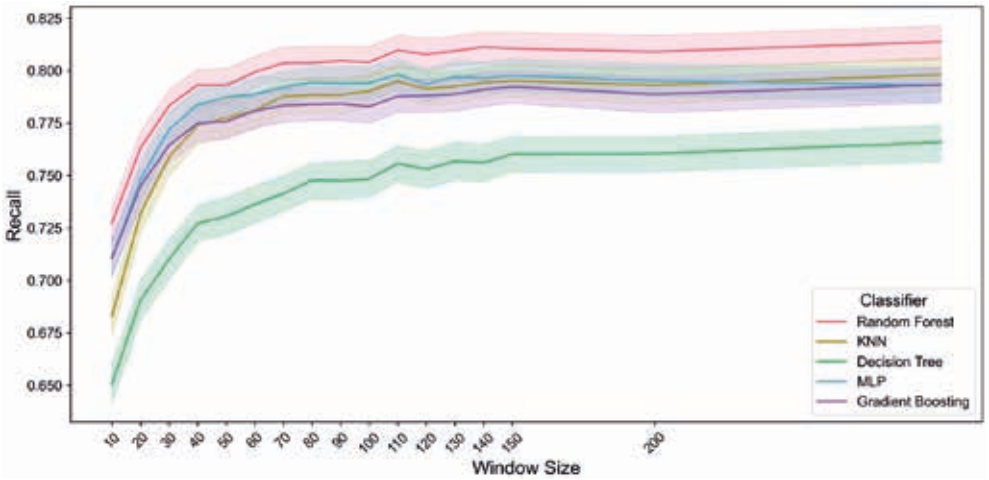


Fig. 9 - Effect of *W* on classification recall. This figure illustrates the variation in classification recall for different ML models as a function of sliding *W*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

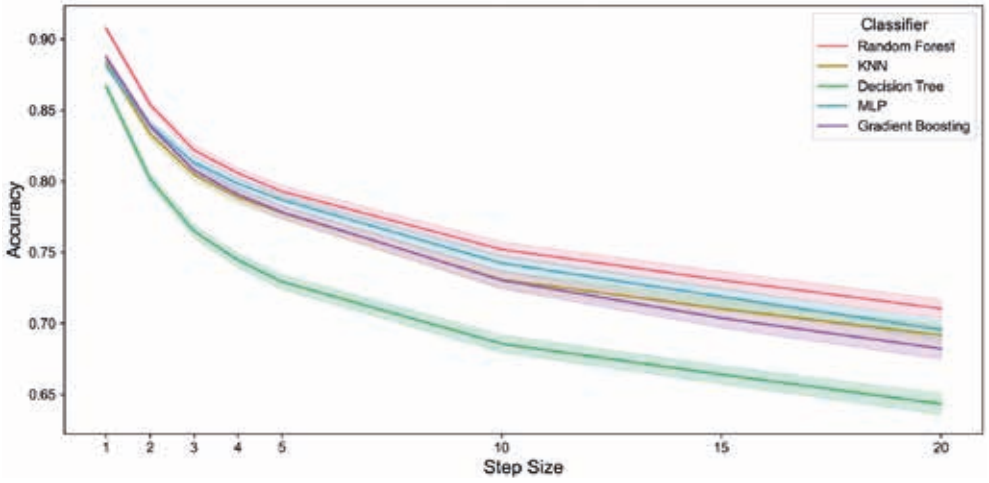


Fig. 10 - Effect of *S* on classification accuracy. This figure illustrates the variation in classification accuracy for different ML models as a function of *S*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

3.2.2. Precision

Precision declines with increasing *S* (Fig. 11). The most rapid decrease occurs between *S* = 1 and *S* = 5, where precision drops from values above 0.80 to approximately 0.72–0.75, after which the curves level off beyond *S* ≈ 10. RF performs best at small steps, while MLP becomes dominant at moderate steps. KNN and DT show substantially lower precision at larger *S*s.

3.2.3. F1-score

The F1-score decreases as *S* increases (Fig. 12). The steepest reduction is observed between *S* = 1 and *S* = 3, with the F1-score decreasing by approximately 6–9%, followed by stabilisation

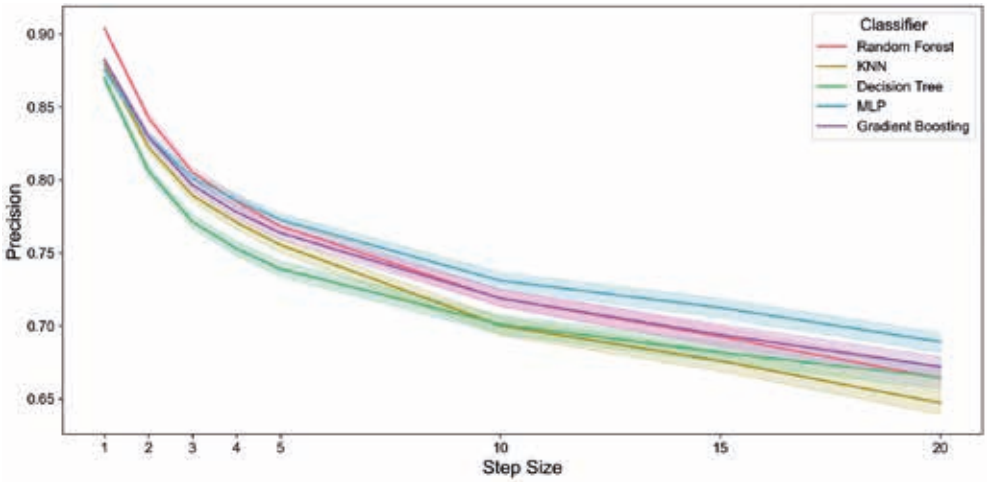


Fig. 11 - Effect of *S* on precision. This figure illustrates the variation in precision for different ML models as a function of *S*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

beyond *S* = 10. RF performs best at very small steps, while MLP becomes dominant at larger steps. DT shows the largest performance degradation.

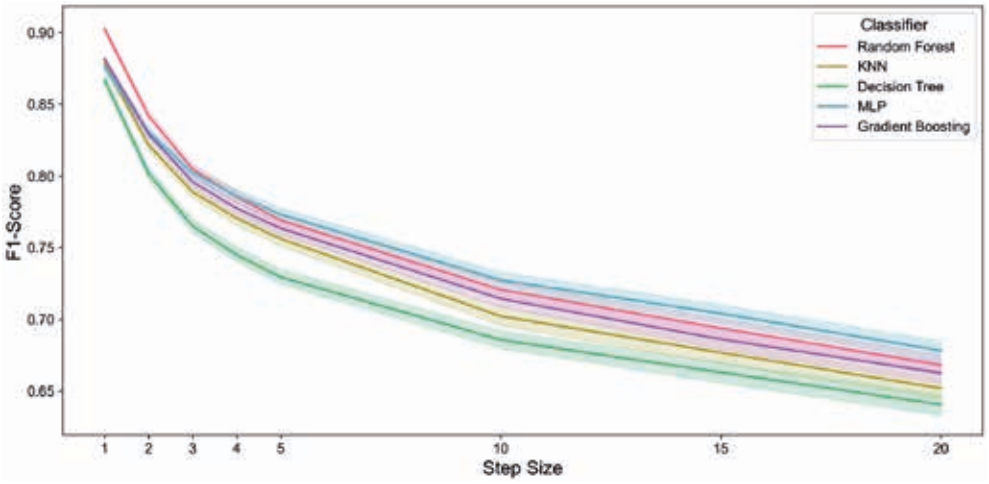


Fig. 12 - Effect of *S* on F1-score. This figure illustrates the variation in F1-score for different ML models as a function of *S*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

3.2.4. Recall

Recall declines steadily with increasing *S* (Fig. 13), with the most significant drop between *S* = 1 and *S* = 3, where recall decreases from values above 0.82 to approximately 0.75. RF maintains the most stable recall, whereas DT experiences the strongest decline.

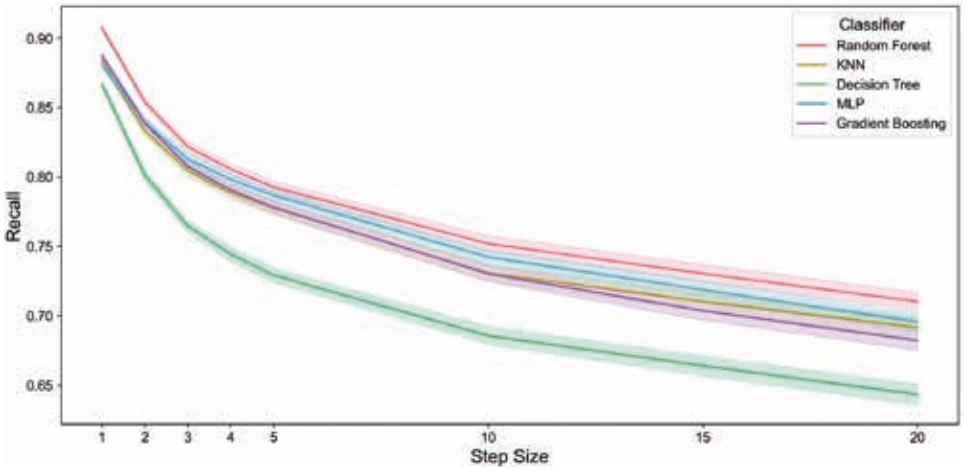


Fig. 13 - Effect of *S* on recall. This figure illustrates the variation in recall for different ML models as a function of *S*. The shaded regions represent confidence intervals, highlighting the stability and generalisation of each model.

3.3. Effect of the informational richness indicator

The effect of the $IRI_{\%}$ on classifier performance is summarised using three regimes: low $IRI_{\%}$ regime (<5%), mid $IRI_{\%}$ regime (5–50%), and high $IRI_{\%}$ regime (>50%). Across all classifiers and metrics, performance improves with increasing $IRI_{\%}$ values and, then, stabilises at higher values. Periodic local dips are observed near transitions between (*W*, *S*) bands, most notably around $IRI_{\%}$ values ≈ 58–62%.

3.3.1. Accuracy

Fig. 14 shows that accuracy increases monotonically with the $IRI_{\%}$. RF achieves the highest accuracy across all $IRI_{\%}$ bands, while DT exhibits the largest variability at a low $IRI_{\%}$. At a high $IRI_{\%}$ (>50%), all classifiers reach a stable accuracy plateau.

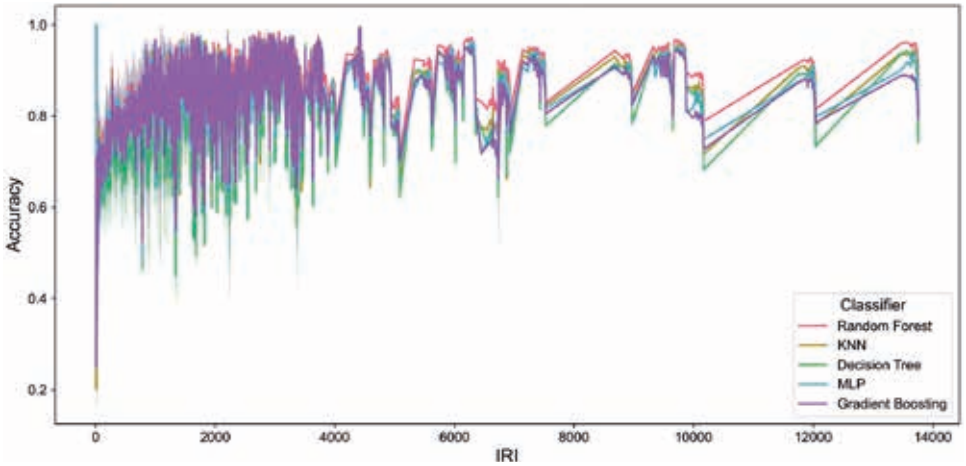


Fig. 14 - This figure illustrates the variation in accuracy as a function of the $IRI_{\%}$: accuracy increases monotonically, shows a small mid-range dip, and levels off high (≈0.9–1.0). RF leads overall; DT is most variable at a low *IRI*.

3.3.2. Precision

Precision as a function of the $IRI_{\%}$ is presented in Fig. 15. Precision improves rapidly from low to mid $IRI_{\%}$ values and saturates at high $IRI_{\%}$ values. RF, MLP, and GB converge toward similar high-precision values in the high-$IRI_{\%}$ regime, while KNN and DT remain lower.

3.3.3. F1-score

The F1-score trends shown in Fig. 16 closely follow those of precision and recall. The F1-score increases steeply from low to mid $IRI_{\%}$ values and, then, flattens beyond $IRI_{\%}$ values ≈ 50%. RF consistently attains the highest F1-scores, followed closely by MLP and GB. DT shows the greatest variability at low $IRI_{\%}$ values.

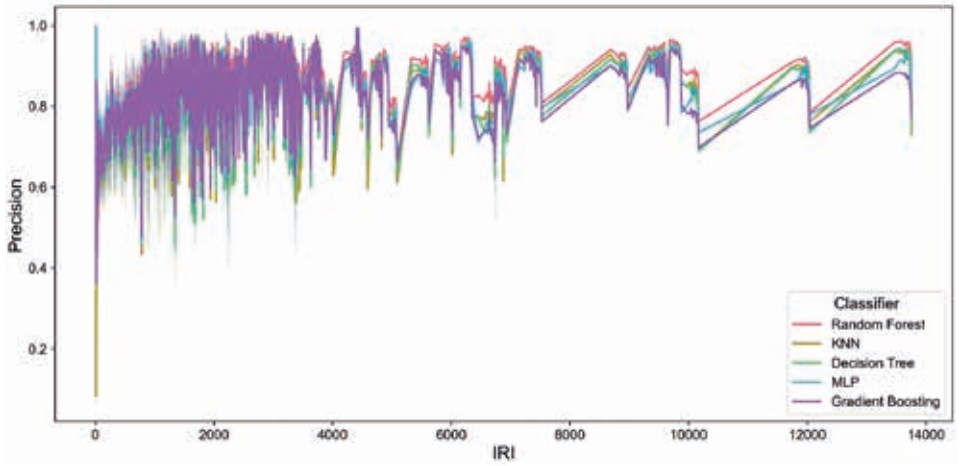


Fig. 15 - This figure illustrates the variation in precision as a function of the $IRI_{\%}$: large spread at low *IRI* values that tighten with greater *IRI* values, with a short boundary dip before stabilising at 0.9. RF and GB dominate; DT shows the widest early spread.

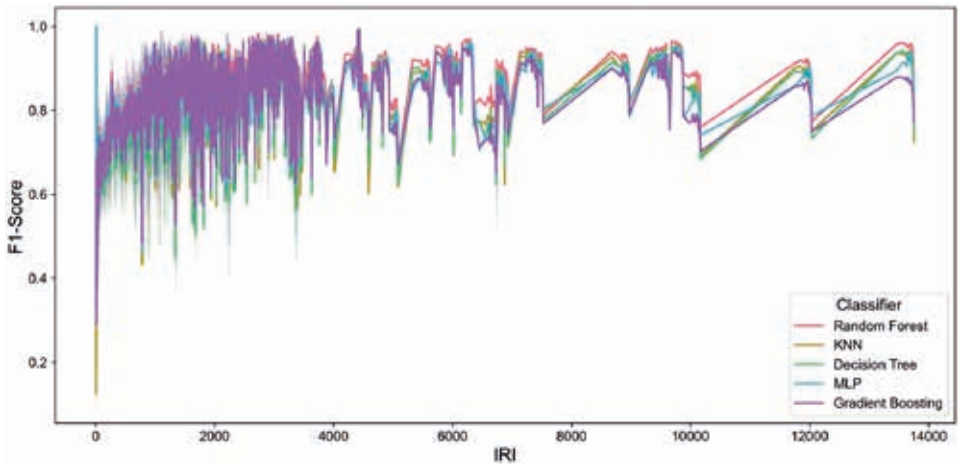


Fig. 16 - This figure illustrates the variation in the F1-score as a function of the $IRI_{\%}$: early fluctuations compress as the *IRI* grows, a transient mid-band dip appears, and performance plateaus near 0.9–1.0. RF remains the safest performer, with GB close behind.

3.3.4. Recall

Recall variations with the $IRI_{\%}$ are shown in Fig. 17. Recall improves rapidly from low to mid $IRI_{\%}$ values and stabilises at high $IRI_{\%}$ values. RF achieves the highest and most stable recall, while DT stabilises at lower recall values and remains the most sensitive to low $IRI_{\%}$ values.

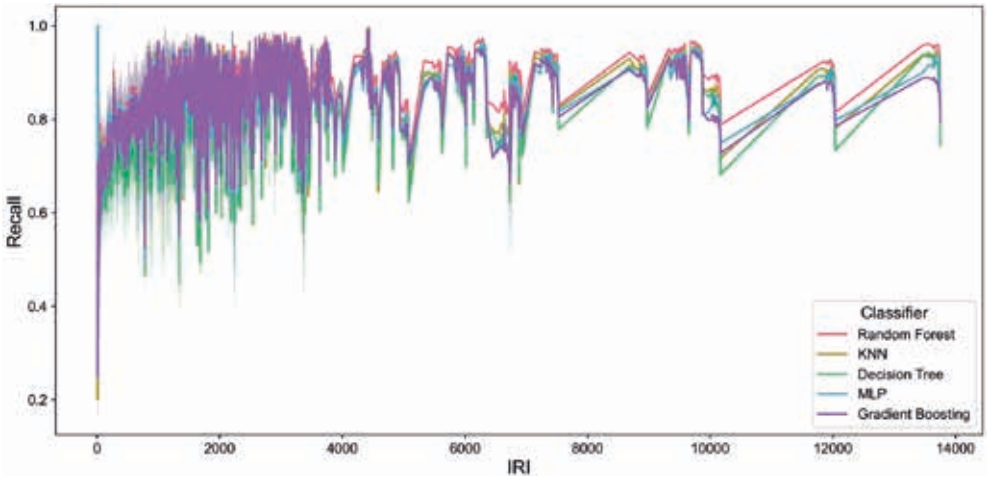


Fig. 17 - This figure illustrates the variation in recall as a function of the $IRI_{\%}$: high volatility at a very low *IRI*, a brief mid-band dip, then, a plateau near 0.9–1.0 at a high *IRI*. RF/GB form the upper envelope; DT and KNN are most unstable early.

3.4. Representative well-level lithology discrimination

To complement the global performance metrics and provide a qualitative assessment of lithology discrimination, a representative well was selected for detailed comparison of predicted lithology profiles obtained using different ML classifiers (Fig. 18). The well was chosen based on data completeness and the presence of multiple lithological transitions, making it suitable for evaluating model behaviour at facies boundaries.

Fig. 18 compares the true lithology log with predictions generated by DT, GB, MLP, RF, and KNN using the same entropy-feature configuration and sliding-window parameters. Visual inspection shows that ensemble-based models, particularly RF, reproduce lithological boundaries more consistently and exhibit fewer spurious class switches. In contrast, simpler models such as DT and KNN display increased fragmentation and misclassification near transition zones.

4. Discussion

This section interprets the experimental results presented above and discusses their implications for lithology discrimination from geophysical well logs, with particular emphasis on the role of sliding-window parameters and the $IRI_{\%}$.



Fig. 18 - Comparison of reference and predicted lithology profiles for a representative well using different classifiers. The true lithology is shown alongside DT, GB, MLP, RF, and KNN predictions computed with identical entropy features and sliding-window settings. RF and GB exhibit the most continuous and stable boundaries, while DT and KNN show higher sensitivity to lithological transitions and greater prediction variability.

4.1. Interpretation of window size effects

Increasing the W improves classification performance up to a threshold (approximately W = 100). Larger windows provide additional contextual information, reducing prediction uncertainty and enhancing the stability of entropy-based descriptors. Small windows (W < 60) offer limited contextual coverage, resulting in higher variability and lower accuracy. Medium windows (W = 60–100) achieve the most favourable balance, preserving local signal variations while providing sufficient contextual richness. For larger windows (W > 100), performance gains plateau, suggesting increasing redundancy that dilutes localised lithological features.

4.2. Interpretation of step size effects

S controls the overlap between consecutive windows and, thus, directly affects data density. Small steps (S = 1–3) generate dense, overlapping segments that preserve fine-scale geophysical variations and yield higher performance. Moderate steps (S = 3–10) lead to gradual performance degradation as overlap decreases. Large steps (S > 10) produce sparse segmentation, significantly reducing classification performance due to the loss of structural detail.

4.3. Interpretation of informational richness indicator effects

The $IRI_{\%}$ provides a quantitative measure of data richness produced by the sliding-window process. Low $IRI_{\%}$ values correspond to sparse segmentation and insufficient contextual information, leading to high variance and reduced recall. Mid-range $IRI_{\%}$ values (5–50%) mark a rapid improvement region, where added overlap introduces meaningful diversity. High $IRI_{\%}$ values (>50%) correspond to a stable performance plateau, beyond which additional overlap yields diminishing returns. Brief performance dips near an $IRI_{\%} \approx$ 58–62% are attributed to transitions between (W, S) parameter bands rather than loss of informational value.

4.4. Classifier performance across parameters

Across all parameter settings, RF demonstrates the greatest robustness and stability, consistently achieving the highest performance in the high-$IRI_{\%}$ regime. GB and MLP closely follow RF, while KNN benefits progressively from richer segmentation but remains less competitive overall. DT is the most sensitive to informational scarcity, exhibiting pronounced performance degradation and variability at a low $IRI_{\%}$ value.

These global trends are further supported by the representative well analysis, which provides a qualitative perspective on classifier behaviour along depth. Ensemble-based models, particularly RF, produce smoother and more geologically consistent lithology profiles, with clearer preservation of stratigraphic boundaries and fewer spurious class switches. In contrast, simpler models such as DT and KNN show increased fragmentation and misclassification near lithological transitions. This qualitative evidence reinforces the quantitative findings and highlights the superior stability of ensemble methods under varying segmentation conditions.

4.5. Parameter interaction and optimisation

The results indicate that optimal performance is achieved by targeting a suitable $IRI_{\%}$ rather than tuning W and S independently. An $IRI_{\%} \geq$ 50% yields stable, near-optimal performance across all metrics. Practical parameter combinations such as W = 60–100 and $S \leq 3$ naturally place the

system in this regime. For computationally constrained scenarios, operating at an $IRI_{\%} \approx 10–20\%$ offers a reasonable trade-off between accuracy and efficiency.

4.6. Practical implications

The findings provide clear guidance for efficient lithology discrimination from well logs. Selecting *W* and *S* based on the $IRI_{\%}$ ensures sufficient informational richness while avoiding unnecessary computational cost. RF emerges as a strong default model for operational deployment, while MLP and GB offer competitive alternatives. Maintaining fine segmentation remains critical for preserving recall and balanced performance, particularly in intervals characterised by subtle or rapidly varying lithological changes.

5. Conclusions

This study proposes a data-richness-driven framework for entropy-based lithology discrimination from geophysical well logs, validated on a large multi-basin dataset comprising 117 wells. Multiple entropy measures were computed under systematic sliding-window configurations and evaluated using representative ML classifiers, enabling a comprehensive assessment of the effects of *W*, *S*, and segmentation density on classification performance.

A key contribution is the introduction of the $IRI_{\%}$, which unifies the combined effects of *W* and *S* into a single quantitative control parameter. Across all classifiers and performance metrics, classification performance increases monotonically with the $IRI_{\%}$ and stabilises beyond approximately 50%, indicating a regime of sufficient informational density. Below this threshold, sparse segmentation leads to higher variance and reduced recall, particularly for simpler models.

Ensemble-based classifiers, especially RF, consistently demonstrate the greatest robustness across segmentation regimes, while GB and MLP show comparable performance at high $IRI_{\%}$ values. KNN benefits progressively from richer segmentation, whereas DT remains the most sensitive to informational scarcity. Practical parameterisation shows that *W*s of 60–100 samples combined with *S*s ≤ 3 reliably achieve $IRI_{\%}$ values ≥ 50%, offering an effective balance between predictive accuracy and computational efficiency.

Overall, centring segmentation design and model tuning on the $IRI_{\%}$ provides a principled and scalable framework for optimising entropy-based lithology classification. This approach enables stable, high-confidence predictions while avoiding unnecessary computational cost, and offers a practical pathway towards standardised deployment in real-world reservoir characterisation workflows.

Acknowledgments. The authors would like to acknowledge the developers of NeuroKit2, an open-source Python toolbox for neurophysiological signal processing. Its robust and versatile framework supported the analysis of geophysical well-log data in this study. We are grateful to the NeuroKit2 community for their ongoing contributions and commitment to open science.

REFERENCES

Amezquita-Sanchez J.P.; 2021: *Entropy algorithms for detecting incipient damage in high-rise buildings subjected to dynamic vibrations*. J. Vib. Control, 27, 426-436, doi: 10.1177/1077546320929145.

Amigó J.M., Balogh S.G. and Hernández S.; 2018: *A brief review of generalized entropies.* Entropy, 20, 813, doi: 10.3390/e20110813.

Bandt C. and Pompe B.; 2002: *Permutation entropy: a natural complexity measure for time series.* Phys. Rev. Lett., 88, doi: 10.1103/PhysRevLett.88.174102.

Borowska M.; 2015: *Entropy-based algorithms in the analysis of biomedical signals.* Stud. Logic Gramm. Rhetor., 43, 21-32, doi: 10.1515/slgr-2015-0039.

Bosch M., Zamora M. and Utama W.; 2002: *Lithology discrimination from physical rock properties*. Geophysics, 67, 573-581, doi: 10.1190/1.1468618.

Boulassel A., Zaourar N., Gaci S. and Boudella A.; 2021: *A new multifractal analysis-based approach for identifying the reservoir fluid nature.* J. Appl. Geophys., 185, doi: 10.1016/j.jappgeo.2020.104185.

Boulassel A., Makhlouf S., Cheddad F.A., Boumelit Z., Zegagh B., Boufenchouche S., Boudella A., Zaourar N. and Gaci S.; 2025a: *Logging-data-driven fluid prediction in clastic reservoir based on fractal attributes and machine learning methods*. In: Gaci S. (ed), Geophysical Exploration for Hydrocarbon Reservoirs, Geothermal Energy, and Carbon Storage: New Technologies and AI-Based Approaches, Wiley, pp. 303-322, doi: 10.1002/9781394261567.ch15.

Boulassel A., Makhlouf S., Boumelit Z., Zegagh B., Boufenchouche S., Cheddad F.A., Boudella A., Zaourar N. and Gaci S.; 2025b: *Unlocking deeper insights: using machine learning to predict dynamic shear wave slowness from well logs.* In: Gaci S. (ed), Geophysical Exploration for Hydrocarbon Reservoirs, Geothermal Energy, and Carbon Storage: New Technologies and AI-Based Approaches, Wiley, pp. 323-342, doi: 10.1002/9781394261567.ch16.

Boulassel A., Makhlouf S., Cheddad F.A., Boumelit Z. and Boudella A.; 2025c: *Reconstructing missing dynamic shear wave slowness from geophysical well logs using ensemble-based machine learning methods*. In: International Conference on Fossil and Renewable Energies (ICFRE' 2025), Boumerdes, Algeria.

Boulassel A., Makhlouf S., Cheddad F.A. and Boumelit Z.; 2025d: *Robust Oil-Gas Discrimination from well logs using wavelet-based multifractal attributes*. In: International Conference on Fossil and Renewable Energies (ICFRE' 2025), Boumerdes, Algeria, November 18-19, 2025.

Boulassel A., Makhlouf S., Boumelit Z., Boudella A., Cheddad F.A., Gaci S., Akli A., Kebir M., Boufenchouche S., Zegagh B., Tedjani A. and Reghis A.; 2026: *Fractal and multifractal analysis of well logs coupled with PLS-DA for fluid characterization in petroleum reservoirs.* Acta Geophysica, 74(2), 119, doi: 10.1007/s11600-026-01824-5.

Breiman L.; 2001: *Random forests*. Machine Learning, 45, 5-32, doi: 10.1023/A:1010933404324.

Cheddad F.A., Makhlouf S., Boulassel A. and Boumelit Z.; 2025: *Saturation prediction from well logs using long short-term memory networks (LSTM): a field case study.* In: International Conference on Fossil and Renewable Energies (ICFRE' 2025), Boumerdes, Algeria, November 18-19, 2025.

Clausius R.; 1865: *Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie.* Ann. Phys., 201, 353-400, doi: 10.1002/andp.18652010702.

Cover T. and Hart P.; 1967: *Nearest neighbor pattern classification*. IEEE Trans. Inf. Theory, 13, 21-27.

De Luca A. and Termini S.; 1972: *A definition of non-probabilistic entropy in the setting of fuzzy sets theory.* Inform. Control, 20, 301-312, doi: 10.1016/S0019-9958(72)90199-4.

Fagbemi O.I., Olayinka A.I., Oladunjoye M.A. and Edigbue P.I.; 2024: *Focused reservoir characterization: analysis of selected sand units using well log and 3-D seismic data in 'Kukih' field, Onshore Niger Delta, Nigeria*. Sci. Rep., 14, doi: 10.1038/s41598-024-56100-7.

Fang Z., Ba J., Guo Q. and Xiong F.; 2024: *Shear-wave velocity prediction of tight reservoirs based on poroelasticity theory: a comparative study of deep neural network and rock physics model*. Geoenergy Sci. Eng., 240, doi: 10.1016/j.geoen.2024.213028.

Ferreira R.B., Vieira V.M., Gleria I. and Lyra M.L.; 2009: *Correlation and complexity analysis of well logs via Lyapunov, Hurst, Lempel–Ziv and neural network algorithms.* Physica A, 388(5), 747-754, doi: 10.1016/j.physa.2008.11.002.

Figiel M. and Lewandowska-Smierzchalska J.; 2022: *Identification of strata from irregularities in well logs using chaos quantifiers.* SPE J., 27, 1770-1781, doi: 10.2118/209236-PA.

Friedman J.H.; 2001: *Greedy function approximation: a gradient boosting machine.* Ann. Statist., 29, 1189-1232, doi: 10.1214/aos/1013203451.

Gilbert R., Liu Y., Abriel W. and Preece R.; 2004: *Reservoir modeling: Integrating various data at appropriate scales*. The Leading Edge, 23, 784-788.

Goodfellow I., Bengio Y. and Courville A.; 2016: *Deep learning*. MIT Press, Cambridge, MA, 800 pp.

Hernandez-Martinez E., Perez-Muñoz T., Velasco-Hernandez J.X., Altamira-Areyan A. and Velasquillo-Martinez L.; 2013: *Facies recognition using multifractal Hurst analysis: applications to well-log data*. Math. Geosci., 45, 471-486, doi: 10.1007/s11004-013-9445-6.

Hristopulos D.T. and Varouchakis E.A.; 2023: *Maximum entropy method*. In: Sagar D., Cheng Q., McKinley J. and Agterberg F. (eds), Encyclopedia of Mathematical Geosciences, Springer International Publishing, Cham, pp. 842-845.

Huang Y., Zhao Y., Capstick A., Palermo F., Haddadi H. and Barnaghi P.; 2024: *Analyzing entropy features in time-series data for pattern recognition in neurological conditions*. Artif. Intell. Med., 150, doi: 10.1016/j.artmed.2024.102821.

Imran Q.S., Siddiqui N.A., Latif A.H.A., Bashir Y., Ali A.A.A.S. and Jamil M.; 2020: *Integrated well data and 3D seismic inversion study for reservoir delineation and description.* Bull. Geol. Soc. Malaysia, 70, 209-220, doi: 10.7186/bgsm70202016.

Kim D., Choi J., Kim D. and Byun J.; 2020: *Predicting mineralogy by integrating core and well log data using a deep neural network.* J. Pet. Sci. Eng., 195, doi: 10.1016/j.petrol.2020.107838.

Lai J., Su Y., Xiao L., Zhao F., Bai T., Li Y., Li H., Huang Y., Wang G. and Qin Z.; 2024: *Application of geophysical well logs in solving geologic issues: past, present and future prospect.* Geosci. Front., 15, doi: 10.1016/j.gsf.2024.101779.

Li Z., Deng S., Hong Y., Wei Z. and Cai L.; 2024: *A novel hybrid CNN–SVM method for lithology identification in shale reservoirs based on logging measurements.* J. Appl. Geophys., 223, doi: 10.1016/j.jappgeo.2024.105346.

Mageed I.A. and Zhang Q.; 2022: *An introductory survey of entropy applications to information theory, queuing theory, engineering, computer science, and statistical mechanics.* In: Proceedings 27th International Conference on Automation and Computing (ICAC), IEEE, 1-6.

Makowski D., Pham T., Lau Z.J., Brammer J.C., Lespinasse F., Pham H., Schölzel C. and Chen S.A.; 2021: *NeuroKit2: a Python toolbox for neurophysiological signal processing.* Behav. Res. Methods, 53, 1689-1696, doi: 10.3758/s13428-020-01516-y.

Mikes D., Barzandji O.H.M., Bruining J. and Geel C.; 2006: *Upscaling of small-scale heterogeneities to flow units for reservoir modelling.* Mar. Pet. Geol., 23, 931-942, doi: 10.1016/j.marpetgeo.2005.06.005.

Ministry of Hydrocarbons and Mines; n.d.: *Hydrocarbons: Sedimentary basins and petroleum provinces of Algeria.* Government of Algeria. https://www.oilmines.gov.dz/?rubrique=hydrocarbures, accessed 26 August 2026.

Mohseni A., Duchaine V. and Wong T.; 2024: *Improvement in Monte Carlo localization using information theory and statistical approaches*. Eng. Appl. Artif. Intell., 131, doi: 10.1016/j.engappai.2024.107897.

Noroozi M. and Riahi M.A.; 2025: *Using deep learning for facies classification in geological exploration: feedforward neural network vs. convolutional neural network approaches.* Bull. Geoph. Ocean., 66, 269–280, doi: 10.4430/bgo00495.

Pincus S.M.; 1991: *Approximate entropy as a measure of system complexity.* Proc. Natl. Acad. Sci., 88, 2297-2301, doi: 10.1073/pnas.88.6.2297.

Powers D.M.; 2020: *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* arXiv:2010.16061, doi: 10.48550/arXiv.2010.16061.

Quinlan J.R.; 1986: *Induction of decision trees.* Machine Learning, 1, 81-106, doi: 10.1007/BF00116251.

Ren X., Hou J., Song S., Liu Y., Chen D., Wang X. and Dou L.; 2019: *Lithology identification using well logs: a method by integrating artificial neural networks and sedimentary patterns.* J. Pet. Sci. Eng., 182, doi: 10.1016/j.petrol.2019.106336.

Roberts S.J., Penny W. and Rezek I.; 1999: *Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing.* Med. Biol. Eng. Comput., 37, 93-98.

Rosenblatt F.; 1958: *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychol. Rev., 65, 386-408, doi: 10.1037/h0042519.

Sokolova M. and Lapalme G.; 2009: *A systematic analysis of performance measures for classification tasks.* Inform. Process. Manage., 45, 427-437, doi: 10.1016/j.ipm.2009.03.002.

Strydom T., Dalla Riva G.V. and Poisot T.; 2021: *SVD Entropy reveals the high complexity of ecological networks.* Front. Ecol. Evol., 9, doi: 10.3389/fevo.2021.623141.

Sun Z., Jiang B., Li X., Li J. and Xiao K.; 2020: *A data-driven approach for lithology identification based on parameter-optimized ensemble learning*. Energies, 13, doi: 10.3390/en13153903.

Wang H. and Chen S.; 2023: *Insights into the application of machine learning in reservoir engineering: current developments and future trends.* Energies, 16, doi: 10.3390/en16031392.

Yang L. and Sun S.Z.; 2020: *Seismic horizon tracking using a deep convolutional neural network*. J. Pet. Sci. Eng., 187, doi: 10.1016/j.petrol.2019.106709.

Yu S. and Ma J.; 2021: *Deep learning for geophysics: current and future trends*. Rev. Geophys., 59, doi: 10.1029/2021RG000742.

Zegagh B., Eladj S., Mihoubi A., Boufenchouche S., Boulassel A., Gaci S., Farfour M., Kehila A., Arab M. and Bennamane K.; 2025: *Lateral lithologic distribution of Triassic formations in the Algerian Oued Mya basin using combined petro-elastic and seismic inversion attributes.* Bull. Geoph. Ocean, 66, 297-316, doi: 10.4430/bgo00489.

Zeroug S., Bounoua N. and Lounissi R. (eds); 2007: *Well Evaluation Conference Algeria 2007.* Schlumberger, Houston, TX, USA, 489 pp.

Zhang P., Liang X., Xian C., Liu B., Wang W. and Zhang C.; 2022: *Geomechanics simulation of stress regime change in hydraulic fracturing: a case study.* Geomech. Geophys. Geo-Energ. Geo-Resour., 8, 83, doi: 10.1007/s40948-022-00391-5.

Zhang T., Han Z., Chen X. and Chen W.; 2021: *Quantifying randomness and complexity of a signal via maximum fuzzy membership difference entropy.* Measurement, 174, doi: 10.1016/j.measurement.2021.109053.

Corresponding author: Fethi Ali Cheddad
Geophysics Department, University of Sciences and Technology Houari Boumediene
BP 32, El Alia, 16111 Bab Ezzouar, Algiers, Algeria
Phone: +213 671213931; e-mail: fcheddad@usthb.dz