scientific reports

Check for updates

OPEN

Data-driven total organic carbon prediction using feature selection methods incorporated in an automated machine learning framework

Bruno da Silva Macêdo[1], Dennis Delali Kwesi Wayo[2,3], Deivid Campos[4], Rodrigo Barbosa De Santis[5], Alfeu Dias Martinho[6]✉, Zaher Mundher Yaseen[7], Camila Martins Saporetti[8] & Leonardo Goliatt[9]

An accurate assessment of shale gas resources is highly important for the sustainable development of these energy resources. Total organic carbon (TOC) analysis thus becomes fundamental for understanding the distribution and quality of hydrocarbon source rocks within a shale gas reservoir. The elevation of the TOC is often associated with the presence of source rocks, indicating the potential for oil and gas production. TOC assessment is performed using laboratory methods, which can be time-consuming and costly. Data-driven models have been successfully applied to model the relationship between TOC and other constituents and to predict the TOC content. However, these methods depend on extensive parameter adjustments that must be carefully conducted in different sedimentary environments. In this context, Automated Machine Learning (AutoML) is an alternative for accurately predicting TOCs, saving time-consuming fine-tuning steps in model development. This study aims to develop an AutoML strategy for estimating TOC using well log data. This procedure automatically preprocesses the search for the best method parameters, reducing the execution time. Among the methods evaluated, Extremely Randomized Trees (XT) performed best (R = 0.8632, MSE = 0.1806) in the test set. The proposed strategy provides a powerful data-driven method, which allows real-world use of the well to assist in data analysis and subsequent decision-making.

Shale gas has emerged as a significant unconventional resource in the oil and gas industry, driven by advancements in exploration technologies such as horizontal drilling and hydraulic fracturing[1,2]. Accurate assessment of shale gas resources is crucial for sustainable energy development, and in this context, Total Organic Carbon (TOC) content is a critical geochemical parameter[3]. TOC represents the amount of organic matter within rocks and is a decisive indicator of the hydrocarbon generation potential of shale formations. Therefore, TOC analysis is fundamental to understanding the distribution and quality of hydrocarbon source rocks in shale gas reservoirs.

TOC is strongly associated with the presence of source rocks, which are essential for hydrocarbon production[4]. Furthermore, TOC affects critical properties of shale rocks, including porosity, permeability, brittleness, wettability, and diffusivity, all of which are integral to determining the viability and efficiency of shale gas production[5]. Traditionally, the measurement of TOC involves laboratory-based methods, such as wet

[1]Department of Computer Science, Federal University of Lavras, Lavras, MG 37200-000, Brazil. [2]Faculty of Chemical and Process Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, 26300 Kuantan, Malaysia. [3]Department of Petroleum Engineering, School of Mining and Geosciences, Nazarbayev University, 010000 Astana, Kazakhstan. [4]Computational Modeling Program, Engineering Faculty, Federal University of Juiz de Fora, Juiz de Fora 36036-900, Brazil. [5]Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora 36036-900, Brazil. [6]Exact Sciences and Technology Department, Púnguè University, Tete Delegation, Campus Universitário de Cambinde - EN106, Matundo, Tete, Mozambique. [7]Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia. [8]Department of Computational Modeling, Polytechnic Institute, Rio de Janeiro State University, Nova Friburgo 22000-900, Brazil. [9]Department of Computational and Applied Mechanics, Federal University of Juiz de Fora, Juiz de Fora 36036-900, Brazil. ✉email: alfeu.martinho@unipungue.ac.mz

oxidation and dry combustion, which can be both time-consuming and costly[6,7]. Rocks with a TOC weight percentage above 4% are generally considered high-quality hydrocarbon source rocks[8,9]. Studies have also explored the use of conventional petrophysical log data to estimate TOC, offering an alternative to traditional laboratory methods[10–15].

Core sampling has long been used in oil and gas exploration to gather valuable data on TOC. However, the operational complexities and risks associated with core sampling in production settings, including the potential for wellbore instability, formation damage, and high costs, highlight the need for more efficient methods. The risks inherent in traditional core sampling operations, such as wellbore collapse, lost circulation, and the potential for sample degradation, make the exploration of alternative methods, such as data-driven approaches, highly appealing[16–25]. These challenges are further compounded by the logistics involved in accessing exploration and production fields, which often require specialized equipment and expertise[17].

Data-driven approaches, particularly Machine Learning (ML) models, offer a promising alternative. When validated and strategically implemented, ML models can establish indirect correlations, enhancing efficiency, reducing costs, and improving the understanding of TOC distribution within reservoirs[14,15]. Exploration wells often encounter harsh environments and significant depths, requiring specialized equipment and posing logistical challenges[16,17]. Maintaining sample integrity during extraction under high pressure and temperature is also crucial for accurate TOC analysis[18]. Furthermore, core sampling operations can destabilize the wellbore, leading to potential safety hazards and operational difficulties[19–21], and, in severe cases, formation damage[22–24]. Core samples provide data from limited points, potentially missing crucial information due to spatial heterogeneity[25].

The inherent risks and complexities of traditional TOC assessments[26] make exploring indirect correlations using readily available well log data an attractive option[27]. A calibrated data-driven model can mitigate operational risks, offering a safer and more efficient approach compared to core sampling[28,29]. These models can provide a more comprehensive understanding of TOC distribution across the reservoir, reducing reliance on limited core sample data[30,31]. Researchers have investigated correlations between TOC and various well log parameters, including density, resistivity, gamma rays, neutrons, and acoustics, to develop more agile and cost-effective TOC estimation methods[32,33]. Data-driven models have been successfully employed to model the relationship between TOC and other constituents, and to predict TOC content[34–38]. These models enhance the accuracy, efficiency, and depth of TOC analysis, contributing to a better understanding of gas and oil resources and enabling more informed decision-making in exploration and production.

ML models can process large volumes of data rapidly[39], predict hydrocarbon-generating potential based on various variables[40], and identify key factors influencing TOC variation[41]. They can also optimize resource allocation and sampling[42]. Ensemble methods and feature selection, used in other geoscientific applications such as lithology classification[43–45], production rate forecasting[46], and drilling data classification[47], further demonstrate the potential of ML in this field. However, the effectiveness of ML models can vary depending on the diagenetic processes of source rocks. Optimal model selection and precise configuration of internal parameters are essential considerations for achieving optimal results. It is also crucial to recognize that ML models are not a one-size-fits-all solution for all sedimentary environments. These methods require careful adjustment of the internal parameters, which play a significant role in the final performance of the model. These parameters, often called hyperparameters, profoundly influence the model's ability to learn relevant data patterns and generalize this knowledge to new situations[48]. In this context, Automated Machine Learning (AutoML) models have emerged as promising and flexible alternatives to address these challenges.

A bibliometric search in the Scopus database using terms like ("geosci*", "automl") and ("automat* AND machine AND learning") yielded only three relevant articles[49–51], indicating a significant research gap in AutoML applications in geosciences, particularly for TOC prediction. While AutoML has been successfully implemented across various domains for automating complex tasks such as model selection and hyperparameter optimization[52], its application to TOC prediction could significantly streamline the predictive modeling process, which traditionally requires domain-specific expertise for machine learning model tuning. The integration of AutoML in geosciences presents an opportunity to automate and optimize these traditionally manual processes.

Recently, AutoML approaches have attracted the attention of oil and gas researchers, which have been combined to build a super learning approach to predict TOC[53]. The data used for this work were from the set of oil shales in the Qingshankou Formation in the Songliao Basin, China. The superlearning model resulted in R^2 = 0.80 and RMSE = 1.16. Ensemble approaches are consistently used as alternative approaches to improve accuracy and produce robust model[54,55]. ML models were also hybridized with metaheuristics to exploit their parametric settings. A successful example is the kerogen estimation through the types of petrophysical well logs through the hydrogen and oxygen indices[56–59].

AutoML models offer a more automated, efficient, and affordable approach to building ML models, allowing the implementation and exploitation of ML solutions to be simpler and more effective[60]. The AutoML approach, although computationally more expensive than standalone models, eliminates the need to manually test multiple models and adjust hyperparameters for each of them, saving valuable time and resources. Additionally, they can be configured to fetch the best combinations of models and parameters efficiently, resulting in improved and more consistent performance across different scenarios. The model presented in this paper aims to contribute to geophysical research and the exploration of unconventional resources by incorporating innovative ML methods to address critical problems in exploration and is highly relevant in production.

This paper contributes to geophysical research and unconventional resource exploration by incorporating innovative ML methods to address critical challenges in exploration and production. The proposed model aims to improve the accuracy and efficiency of TOC forecasting, identify patterns in TOC data, and ultimately support more informed decision-making in the oil and gas industry. This study aims to advance the field of geophysics by bringing an automated machine learning methodology to find the models and the internal parameters of these models to efficiently perform TOC predictions. The specific objectives are:

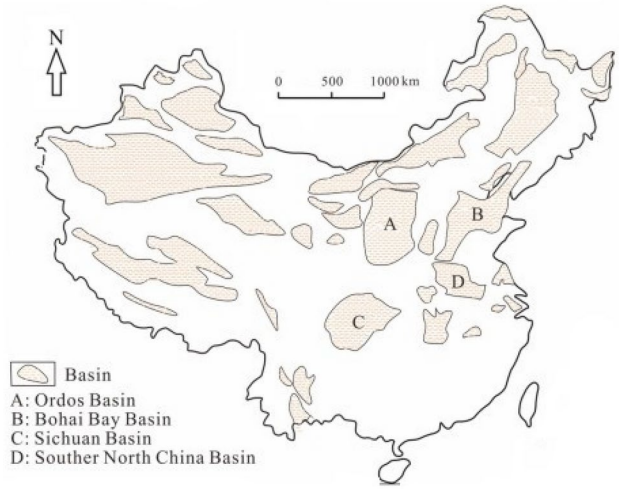

Fig. 1. Distributions of the Ordos Basin, Bohai Bay Basin, Sichuan Basin, and southern North China Basin[5].

	AC	GR	K	RD	RS	TH	U	TOC (%)
Mean	115.328	94.234	0.931	162.957	167.490	6.245	4.994	1.252
Std	76.083	52.712	1.445	138.504	144.280	7.810	4.489	0.797
Min	56.606	32.896	0.000	5.014	5.488	− 0.290	0.140	0.000
Median	66.887	71.174	0.009	133.703	139.226	1.707	3.917	1.102
Max	296.668	450.827	4.622	1002.391	1186.991	39.226	43.056	7.730

Table 1. Statistics related to TOC levels—training set (571 samples).

- To assess the Greedy Weighted Ensemble performance generated by the other AutoML models.
- Incorporate Boruta Feature Selection (BFS), Mutual Information (MI) and Recursive Feature Elimination (RFE) feature selection methods in automated machine learning for TOC prediction.
- Perform a comparative analysis of the predictive performance of different automated machine learning models.

This paper is organized as follows. The following sections describe the dataset and present a detailed automated ML framework proposal. In Section 3, computational experiments are presented, and the comparative performance achieved by ML models used by AutoML is discussed. Finally, Section 4 presents the conclusion.

Materials and methods

Dataset

The present investigation is based on a database available in the literature[5] at https://doi.org/10.3390/en16104159. This database comprises 816 data points related to total organic carbon (TOC) levels and well logs. These data were collected from five shale formations in different geological basins. Among the formations considered are the Yanchang shale in the Ordos Basin and the Shahejie shale in the Bohai Bay Basin. The Longmaxi shale in the Sichuan Basin and the Shanxi and Taiyuan shales are located south of the North China Basin (Fig. 1). These sedimentary formations play crucial roles as hydrocarbon reservoirs, representing paradigmatic cases of oil/gas source rocks in China, with substantial energy resource reserves contained in shale deposits.

A well log is data of the formations and any occurrences found in the well drilling procedure. Such information is useful for evaluating conditions throughout the depth of the well, assisting in analysis and decision-making. The measurements contained in the well logs are collected at discrete depth intervals[5]. This information does not contain the depth of each measurement to prevent an illustration of the characteristics in relation to depth.

Seven commonly used profile parameters were used to consider the borehole profile data. These include measurements of natural gamma radiation (GR), acoustic time difference (AC), deep resistivity (RD), and shallow resistivity (RS), as well as the concentrations of uranium (U), thorium (Th), and potassium (K). These variables were used to identify the most relevant borehole logs for accurately predicting TOC levels.

Tables 1 and 2 provide a detailed analysis of the most relevant statistics associated with total organic carbon (TOC) levels, as well as the borehole logs that were adequately employed in the training and validation process of the model under consideration. In parallel, Fig. 2 provides the correlation matrix corresponding to the carefully selected parameters. The U characteristic has the highest correlation with TOC, followed by GR, but other characteristics, such as AC, have a high correlation with U and GR, influencing TOC prediction. This comprehensive information is essential in promoting a deep insight into the intricate interconnection between

	AC	GR	K	RD	RS	TH	U	TOC (%)
Mean	117.401	97.184	0.928	169.087	173.046	5.725	5.577	1.329
Std	78.630	60.992	1.447	123.712	127.084	7.629	5.350	0.839
Min	56.876	35.083	0.000	4.656	5.084	0.104	0.157	0.040
Median	66.478	70.094	0.008	134.980	143.673	0.495	3.997	1.093
Max	289.349	470.753	4.567	757.350	927.663	25.216	45.424	4.740

Table 2. Statistics related to TOC levels—test set (254 samples).

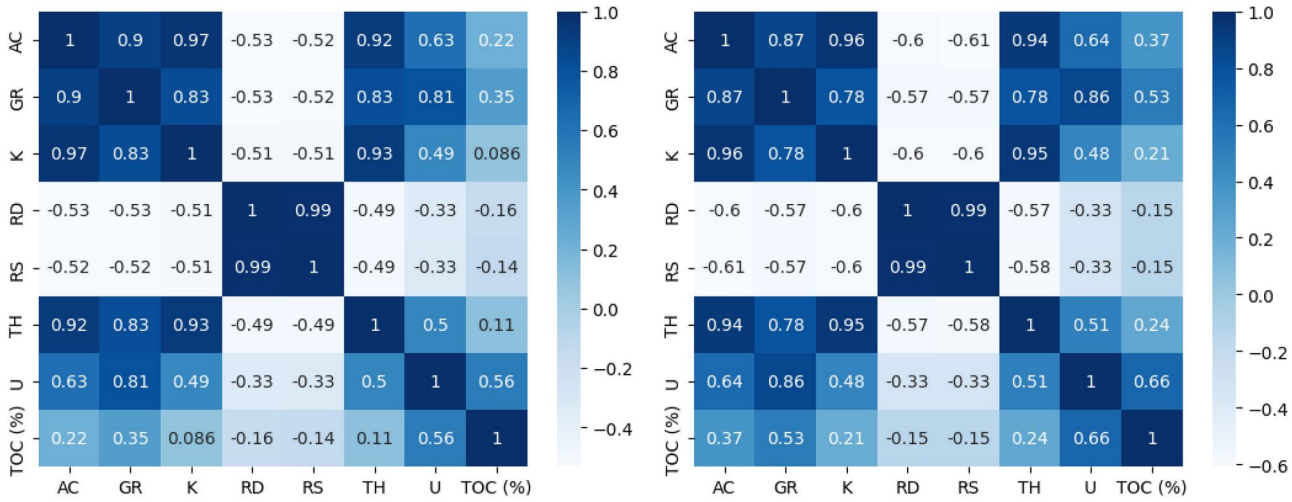

Fig. 2. Correlation coefficients for the training set (left) and test set (right).

fundamental borehole characteristics and associated TOC levels. This deeper understanding not only drives the progress of the investigation but also guides the data split, with a 70% allocation for training and 30% for testing. This process results in a total of 571 samples for training and 245 samples for testing.

Automated machine learning model

AutoML (Automated Machine Learning) is an automated approach for building ML models to simplify and speed up the data modeling process. AutoML aims to reduce the model training time and improve prediction accuracy, allowing users with different levels of ML experience to train high-quality models. AutoML can include multiple steps, such as model selection, data preprocessing, feature selection, hyperparameter tuning, and model evaluation[39].

In the last few years, the impact of AutoML approaches has spanned many industries. In finance, it has been employed to automate the process of selecting the best ML models for different tasks, such as fraud detection, credit scoring, and risk assessment[61]. In the health field, its applications include predicting disease probability based on medical history and genetics[62]. The manufacturing industry benefits from AutoML in predicting equipment failures and optimizing production[63], while in the retail sector, it is used to forecast product demand and improve logistics[64,65].

In optimizing the process of training ML models, adopting AutoML has emerged as an approach of great promise. In this study, an investigation centered on using the automated framework is used to construct and train a predictive model to estimate total organic carbon. The primary objective is to improve predictions while substantially reducing the associated training time.

AutoGluon features a multilayered stacking approach that aims to identify the model with the highest performance and the hyperparameters that make it most effective. At the heart of this process is a meticulous exploration of various ML algorithms, such as neural networks (NNFastAi, NeuralNetTorch), LightGBM-powered trees (LGB)[66], CatBoost-powered trees (CatBoost)[67], Extremely Randomized Trees (XT)[68], Extreme Gradient Boosting (XGBoost)[69], K-Nearest Neighbors (KNN)[70] and Random Forests (RF)[71].

Table 3 shows the hyperparameters used for the automated models. A notable highlight lies in the incorporation of data preprocessing techniques, such as normalization and feature selection, which aim to optimize the accuracy of the resulting model, improve the generalization capacity, and mitigate the risk of overfitting[72]. Figure 3 shows the AutoML steps, from data preparation to hyperparameter optimization and model validation.

The presence of noise in the data is something that can impact the performance of machine learning methods. In this work, we chose to use data normalization in order to smooth out the noise through scale adjustment. There are works that employ methods such as covariance determinant (MCD), stochastic outlier selection (SOS),

Model	Encoding	Description	Range
CatBoost	x_1	Learning rate	$[10^{-4}, 10^{-1}]$
	x_2	Model depth	[1, 16]
	x_3	L_2 regularization	[1, 10]
	x_4	Bagging temp.	[0, 10]
	x_5	Grow policy	['Symmetric', 'Depthwise', 'Lossguide']
Fast Neural Network (NNFastAi)	x_1	No. layers	[1, 5]
	x_2	Dropout prob.	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]
	x_3	Batch size	[16, 512]
	x_4	Learning rate	$[10^{-4}, 10^{-2}]$
NeuralNetTorch (NN_TORCH)	x_1	No. epochs	[5, 50]
	x_2	Learning rate	$[10^{-4}, 10^{-2}]$
	x_3	Activation	['relu', 'softrelu', 'tanh']
	x_4	Dropout prob.	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]
	x_5	Batch size	[16, 512]
	x_6	No. layers	[1, 5]
	x_7	Hidden units	[32, 512]
LightGBM (LGB)	x_1	Learning_rate	[-9, -2]
	x_2	Fature_fraction	[0.1, 1.0]
	x_3	Min_child_samples	[5, 100]
	x_4	Num_leaves	[26, 100]
XGBoost (XGB)	x_1	n_estimators	[10, 100]
	x_2	Learning_rate	$[10^{-4}, 10^{-2}]$
	x_3	Max_depth	[3, 20]
	x_4	Min_child_weight	[1, 20]
	x_5	Colsample_bytree	[0.1, 1.0]
	x_6	Subsample	[0.1, 1.0]
K-Near. Neighb. (KNN)	x_1	n_neighbors	[1, 100]
	x_2	Weights	['uniform', 'distance']
	x_3	Metric	['Euclidean', 'Manhattan', 'Chebyshev']
	x_4	Leaf_size	[10, 100]
	x_5	Exponent p	[1,2]
Random Forest (RF)	x_1	n_estimators	[5, 100]
	x_2	Max_leaf_nodes	[1, 15000]
	x_3	Bootstrap	[True, False]
	x_4	Criterion	[squared_error, absolute_error]
Extreme Trees (XT)	x_1	n_estimators	[5, 100]
	x_2	Max_leaf_nodes	[1, 15000]
	x_3	Bootstrap	[True, False]
	x_4	Criterion	[squared_error, absolute_error]

Table 3. Hyperparameters for automated models.

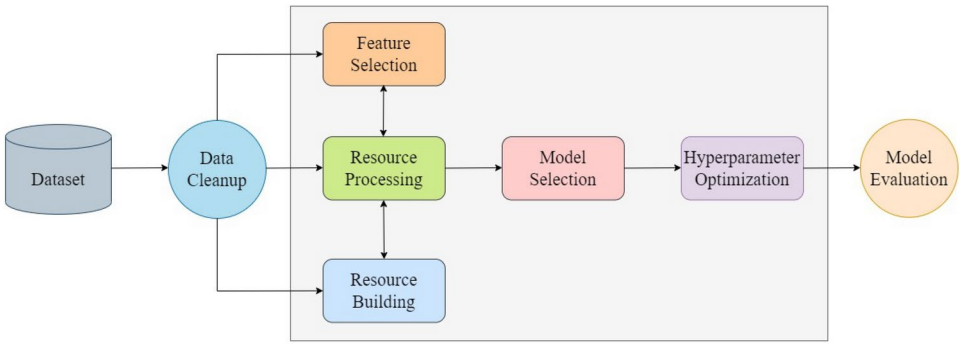

Fig. 3. AutoML model[75].

connectivity-based outlier factor (COF), clustering-based outlier factor (CBOF), Subspace Outlier Detection (SOD), Histogram-based Outlier Score (HBOS), among others[73,74]. Here, we do not apply them because the database contains few samples, and these samples are of different formations, which can drastically reduce the number of samples, which would also impact the performance of the methods.

Feature selection approaches

Feature selection (FS) plays a pivotal role in ML by identifying the most critical features that contribute to model performance and interpretability. FS allows the development of more efficient, accurate, and interpretable ML models and the selection of appropriate feature selection techniques. In this paper, three FS methods were implemented and compared to select the most relevant features.

Boruta feature selection

The Boruta feature selection technique, proposed by[76], is an algorithm with a wrapper approach, using a tree-based ensemble algorithm as the basis estimator. This algorithm was based on some concepts from[77] to define significant input variables by comparing the significance of real resources with that of random probes. The Boruta method has been successfully incorporated as a feature selector in problems in different research areas[78–81].

The Boruta algorithm creates shadow features, which are randomized copies of the original features. These shadow features are then added to the dataset, and a Random Forest classifier is trained on the augmented dataset. For each feature (both original and shadow), the algorithm calculates a Z-score based on the feature's importance, as measured by the Mean Decrease Accuracy (MDA). The MDA is defined as:

$$MDA = \frac{1}{m_{\text{tree}}} \sum_{m_{\text{tree}}}^{m=1} \frac{\sum_{t \in OOB} I(y_t = f(x_t)) - \sum_{t \in OOB} I(y_t = f(x_t^n))}{|OOB|} \tag{1}$$

where m_{tree} represents the number of trees in the tree-based ensemble model, *OOB* denotes the out-of-bag samples, $I(\cdot)$ is the indicator function, y_t is the true target value for sample *t*, $f(x_t)$ is the predicted value before permutation, and $f(x_t^n)$ is the predicted value after permuting feature *n*.

The Z-score for each feature is calculated as:

$$Z\text{-score} = \frac{MDA}{SD} \tag{2}$$

where *SD* represents the standard deviation of the accuracy losses across the trees. The Z-scores of the original features are then compared to the maximum Z-score among the shadow features. Features that consistently achieve significantly higher Z-scores than the shadow features are deemed important and retained, while those with lower or comparable Z-scores are considered irrelevant and discarded. This iterative process continues until all features are either confirmed as important or removed as irrelevant.

Partial mutual information (PMI)

The Mutual Information (MI) feature selection method is a non-linear, multivariate technique rooted in information theory that quantifies the interdependence between variables[82,83]. This method identifies relationships within datasets[84,85]. The PMI score is calculated based on the joint and marginal probability distributions of the variables. Let *U* represent a candidate input variable, *V* the output variable (TOC in this case), and *W* the set of already selected input variables. The PMI between *U* and *V*, conditioned on *W*, is defined as:

$$PMI(U, V \mid W) = \iint f_{U',V'}(u', v') \ln \left[\frac{f_{U',V'}(u', v')}{f_{U'}(u') f_{V'}(v')} \right] dv'\, du' \tag{3}$$

where

$$u' = u - E(u \mid W),\ v' = v - E(v \mid W) \tag{4}$$

represent the residuals of *u* and *v* with respect to *W*, respectively. The functions $f_{U'}(u')$ and $f_{V'}(v')$ denote the marginal probability density functions, and $f_{U',V'}(u', v')$ represents the joint probability density function of u' and v'. For datasets with available samples, the MI score can be estimated using the following equation:

$$PMI(U, V \mid W) = \frac{1}{n} \sum_{i=1}^{n} ln[\frac{f_{U',V'}(u_i', v_i')}{f_{U'}(u_i') f_{V'}(v_i')}] \tag{5}$$

The MI score is symmetric and non-negative. A PMI value of 0 indicates independence between *U* and *V*, given *W*. Higher PMI values signify stronger dependence or association between the variables, suggesting that *U* provides additional information about *V* that is not captured by the variables in *W*. In the context of feature selection, MI allows the identification and selection of features that exhibit high dependence on the target variable while minimizing redundancy with the already selected features. The filtering process selects the top *k* features with the highest MI values as the model's input variables. This process leads to the construction of more efficient and interpretable models with improved predictive performance.

Recursive feature elimination (RFE)

Recursive Feature Elimination (RFE) is a well-established and widely used technique for selecting informative features for model development[86]. This method iteratively selects features from progressively smaller subsets based on an external estimator that assigns weights or importance scores to each feature (e.g., linear model coefficients)[87]. The RFE process starts by training the estimator model on the complete set of features. The importance of each feature is then assessed using a specific attribute measure. Following this assessment, the least important features are progressively eliminated from the original set. This iterative training process, importance evaluation, and feature removal continues until the desired number of features is obtained. As illustrated in Fig. 4, RFE progressively eliminates features with diminishing importance, resulting in a reduced feature set that retains the most informative features for model construction. This has the potential to improve model performance by mitigating the influence of irrelevant features.

Computational experiments

Figure 5 shows the computational framework for the proposed approach. The first process consists of cleaning the data (Data Cleanup), which aims to ensure the quality of the data set for carrying out the subsequent steps. With the completion of cleaning, this data is sent to the Feature Selection stage, which uses the BFS (Boruta Feature Selection), MI (Mutual Information), and RFE (Recursive Feature Elimination) techniques to choose the most relevant attributes. This data, now filtered, is destined for the Resource Processing phase, an important step for preparing the input information for the model. The processed resources are directed to the (Resource Building) stage, the Training Set. This training set will be used in model selection (Model Selection), where different algorithms are tested and adjusted in order to find the most appropriate one to solve the problem under study. This choice is made through hyperparameter optimization (Hyperparameter Optimization), so the model parameters are assigned to their ideal values. The models are evaluated by 5-fold cross-validation, which makes it possible to compare different solutions using a performance measure. The chosen and improved model is applied

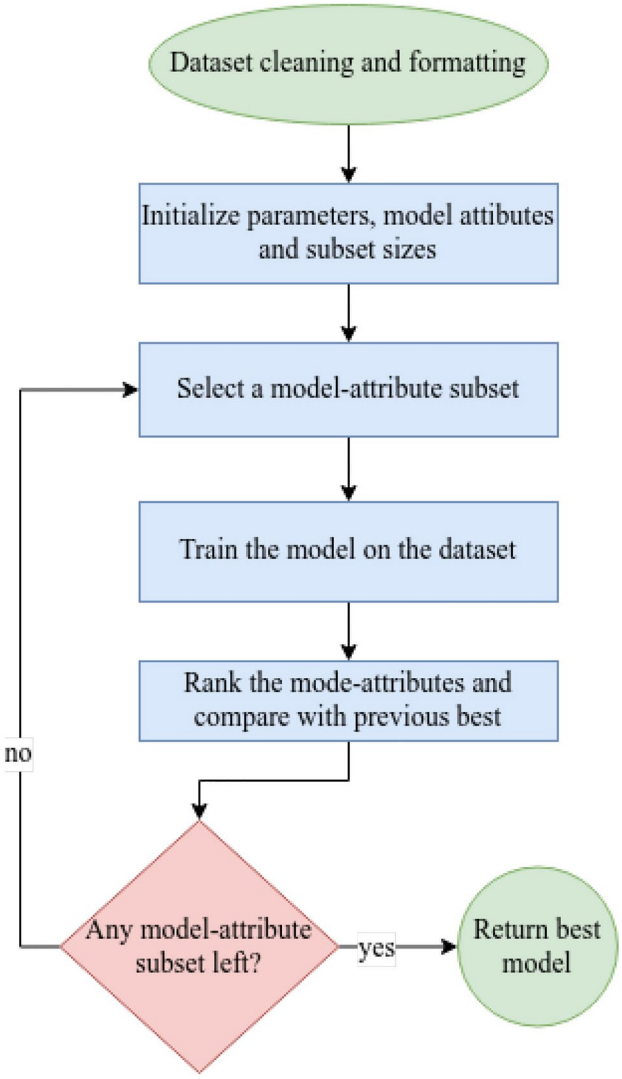

Fig. 4. Workflow for recursive feature elimination[88].

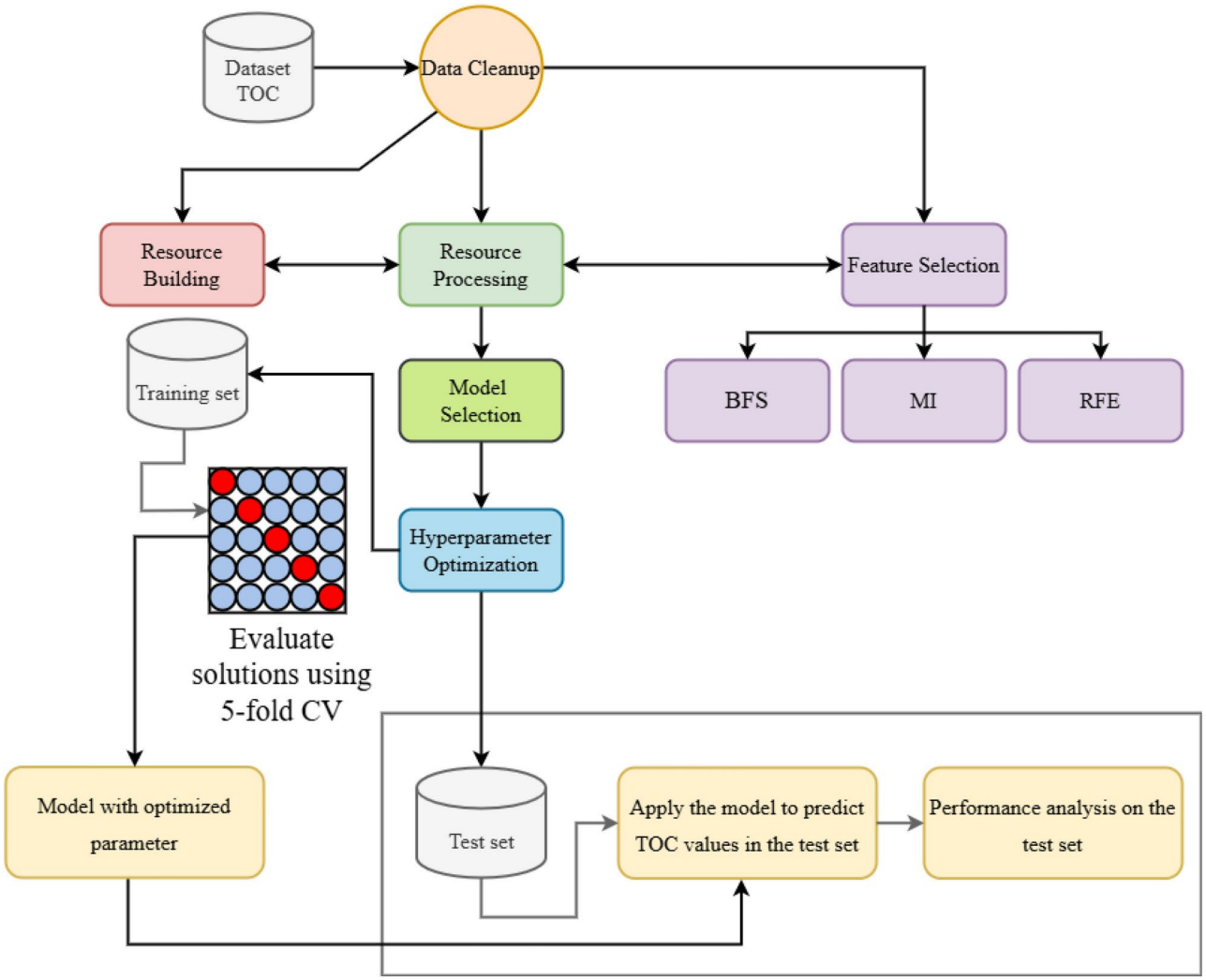

Fig. 5. Framework for the proposed approach.

to the test set (Test Set) to make TOC (Total Organic Carbon) predictions. Finally, the predictions generated by the model are subjected to a Performance Analysis, in which quantitative measurements are employed to evaluate the model's performance on the test set.

A detailed analysis was conducted to identify and correct both missing values and duplicates. By applying automated imputation techniques, missing and duplicate values could be eliminated, simultaneously guaranteeing the coherence and integrity of the underlying data. Data normalization was performed using the z-score standardization technique to make the attributes comparable to scales. This technique ensures that the mean is zero and the variance is one, improving training stability.

AutoGluon uses a grid search approach enhanced by meta-learning techniques, which adapts model selection based on the performance of previous models. This approach speeds up the process and increases the selection accuracy[60]. The k-fold cross-validation technique was applied to evaluate the model's performance and avoid overfitting. Stratified sampling techniques were used to maintain class distribution. This ensures unbiased results and accurate performance estimates. It is known that there is a potential bias generated by the shuffled training and testing divisions, leading each method to predict different test data and be trained on different training data. So this issue was taken into account. 200 independent runs were performed with different seeds, so the grid search employed, which uses Kfold, performed different divisions in the data, allowing a coherent analysis of the results. The models were evaluated using metrics appropriate to the TOC prediction, which is a regression problem. Table 4 shows the metrics used to assess the model's performance.

Table 5 displays the models' evaluation through the metrics RMSE, MSE, MAE, R^2 and R. XT was the method that obtained the best result in the test set, that is, on data not yet seen in the training process by the methods. The results show that some AutoML methods, such as XT and CatBoost, achieve good performance on the test set (data not seen by the methods) for the TOC prediction task. These findings suggest that these methods can be generalized well to new data.

In order to compare the effects of automated search parameters, a baseline model was implemented using default parameters, and its results were presented in Table 5. The parameter of the baseline XGB model according

Performance metric	Acronym	Mathematical expression
Pearson correlation coefficient	R	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$
Coefficient of determination	R^2	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}$
Mean square error	MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
Root mean square error	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
Mean absolute error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$

Table 4. Performance metrics and their mathematical expression.

Model	RMSE	MSE	MAE	R^2	R	Training time (s)
XT	0.425	0.181	0.250	0.743	0.863	0.357
CatBoost	0.445	0.198	0.276	0.718	0.855	20.653
RF	0.459	0.211	0.249	0.700	0.837	0.423
GWE	0.459	0.211	0.270	0.699	0.836	45.321
LGB	0.463	0.214	0.289	0.695	0.835	2.582
XGB	0.463	0.214	0.270	0.695	0.839	5.495
NNTorch	0.484	0.234	0.271	0.666	0.819	8.086
NNFastAi	0.507	0.257	0.312	0.634	0.801	3.869
KNN	0.539	0.290	0.292	0.586	0.767	0.007
Baseline XGB	0.510	0.261	0.289	0.627	0.793	2.277

Table 5. Model evaluation on the test set. GWE: Greedy Weighted Ensemble

to Table 3 is 100 estimators (x_1), learning rate (x_2) equals 0.1, maximum tree depth equals 10 (x_3), child weight (x_4) set to 1, and subsamples (x_5 and x_6) set to 1.0. As can be observed in Table 5, the models with parameter search and optimization achieved superior results in all metrics. This indicates that the parameter tuning process (whether manual or automatic) improves model performance by allowing it to be adapted to the characteristics of the data. This reinforces the importance of research in automated learning models to develop more robust and efficient strategies for determining model parameters.

Figure 6 shows a diagram of the families of models generated by AutoML, the model name, the score value, and the training time. The family that performed best was Greedy Weighted Ensemble_L2, which was generated through a greedy search in combination with methods to create a robust ensemble method. The parameters obtained in this family had a score equal to 0.2346 and a training time of 44.1 seconds. The two methods that obtained good results after the Greedy Weighted Ensemble were LGB and XT. Table 6 presents the best hyperparameters found during the training and the respective validation scores. Although GWE presents the best result in the set, this is not maintained in the test set, as shown in Table 5, likely because it is generated using a greedy search of the training and validation data, which can lead to overfitting.

Table 7 and Fig. 7 show the results incorporating feature selection approaches Boruta, Mutual Information (MI), and Recursive Feature Selection (RFE) into the developed models, considering 5 variables to Boruta, 3 to MI, and 3 to RFE. The assessment is based on root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), coefficient of determination (R^2), and Pearson correlation coefficient (R). Additionally, the training time required for each model is also considered.

Across all the evaluated models, the incorporation of the Boruta feature selection method consistently yielded superior or comparable performance compared to MI and RFE. This is attributed to Boruta's ability to effectively identify the most relevant features for TOC prediction while considering both their individual importance and interdependencies. The inclusion of a larger number of features, as selected by Boruta, provided the models with a more comprehensive representation of the underlying geological and petrophysical relationships influencing the TOC content.

In contrast, MI and RFE, with their tendency to select a smaller subset of features, resulted in models with reduced predictive accuracy. MI, based on information-theoretic principles, can not consider features with complex or non-linear relationships to TOC, while the iterative elimination process that is the basis of RFE can discard features that may contribute to the overall predictive power of the model. Consequently, models

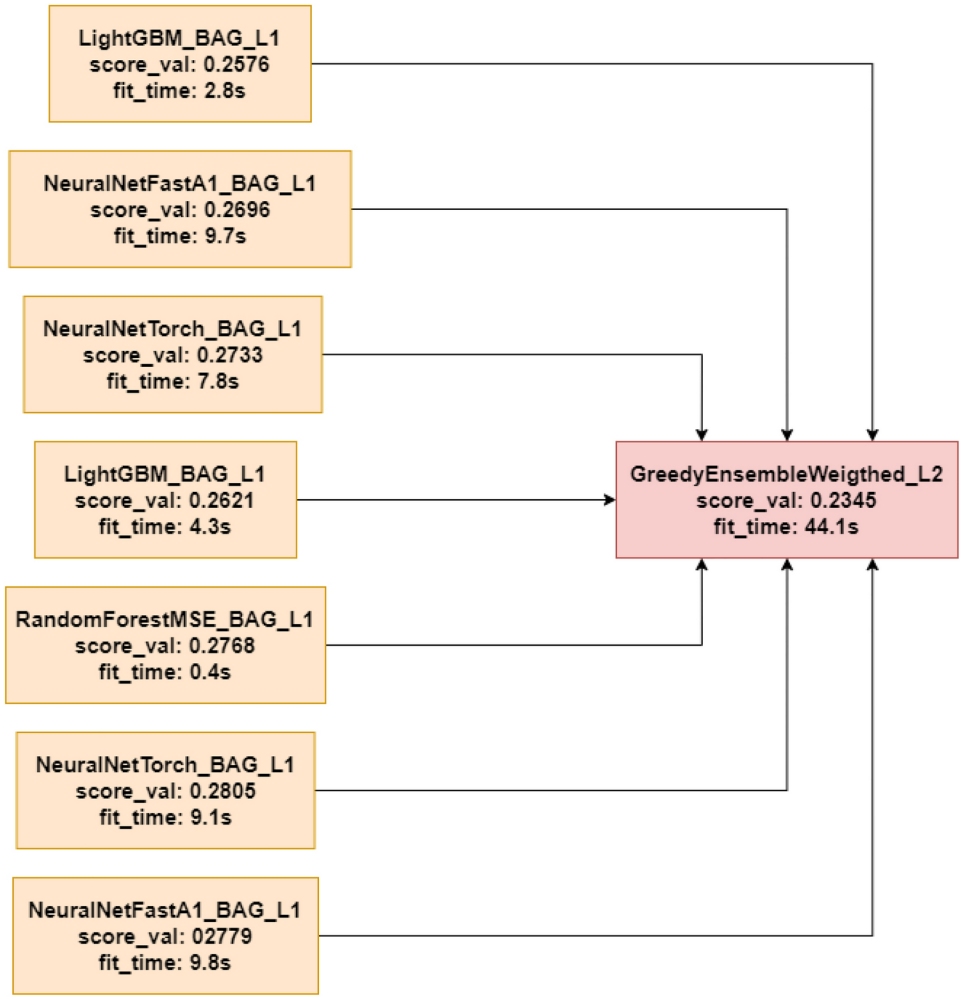

Fig. 6. Greedy Weighted Ensemble model generated by AutoML.

Model	Validation score (MSE)	Best hyperparameters
CatBoost	0.285485	x_1: 0.06027204644225115, x_2: 5, x_3: 3.9883178805569366, x_4: Lossguide
KNN	0.348611	'weights': 'distance'
LGB	0.257551	x_1: 0.042769004147665074, x_2: 0.9988248919194719, x_3: 16, x_4: 31
NNFastAi	0.269638	x_1: (1000, 500, 200), x_2: 0.15108500596256647, x_3: 2048, x_4: 0.00021969615035282952
NeuralNet Torch	0.273252	x_1: 500, x_2: 0.0006353732345430946, x_3: 'tanh', x_4: 0.1, x_5: 512, x_6: 4, x_7: 256
RF	0.276757	x_1: 300, x_2: 15000, x_3: True, x_4: 'squared error'
XGBoost	0.290761	x_1: 10000, x_2: 0.12034886497897113, x_3: 10, x_4: 4, x_5: 0.9871293319286164
XT	0.269325	x_1: 300, x_2: 15000, x_3: True, x_4: 'squared error'
Greedy Weighted Ensemble	0.234488	'ensemble size': 100

Table 6. Best models.

incorporating MI and RFE exhibited higher error values and lower correlation coefficients compared to those utilizing Boruta.

Furthermore, the XT (Extremely Randomized Trees) model emerged as a top performer across all feature selection methods, demonstrating low error values and high correlation coefficients. The robustness and effectiveness of the XT model in handling complex datasets with potentially non-linear relationships make it well-suited for TOC prediction tasks. CatBoost also exhibited strong performance, particularly with the Boruta feature set, highlighting the effectiveness of gradient boosting techniques in this domain.

Examining the final column of the table reveals that model XT achieved the fastest training times, with execution durations of 0.355 and 0.344 seconds, respectively. When considering recursive feature elimination (RFE) for feature selection, both models GWE and XT demonstrated promising results, with training times

FS	Model	RMSE	MSE	MAE	R^2	R	Training time (s)
Boruta (5 var.)	CatBoost	0.450	0.202	0.278	0.712	0.848	8.885
	GWE	0.466	0.217	0.290	0.691	0.832	57.980
	KNN	0.587	0.345	0.343	0.508	0.718	0.076
	LGB	0.484	0.234	0.308	0.667	0.818	3.726
	NNFastAi	0.518	0.268	0.321	0.617	0.787	4.202
	NNTorch	0.477	0.228	0.302	0.675	0.824	30.469
	RF	0.471	0.222	0.259	0.684	0.827	0.401
	XGB	0.485	0.236	0.287	0.664	0.823	4.868
	XT	0.430	0.185	0.264	0.737	0.860	0.355
MI (3 var.)	CatBoost	0.601	0.361	0.336	0.486	0.698	7.472
	GWE	0.576	0.332	0.319	0.527	0.731	47.131
	KNN	0.595	0.354	0.326	0.495	0.709	0.117
	LGB	0.605	0.366	0.329	0.478	0.692	2.621
	NNFastAi	0.635	0.403	0.361	0.426	0.656	7.546
	NNTorch	0.597	0.357	0.346	0.492	0.705	21.821
	RF	0.596	0.356	0.308	0.493	0.709	0.535
	XGB	0.600	0.360	0.335	0.487	0.705	2.159
	XT	0.556	0.309	0.293	0.560	0.748	0.344
RFE (3 var.)	CatBoost	0.518	0.268	0.329	0.618	0.788	2.722
	GWE	0.507	0.257	0.316	0.633	0.797	30.372
	KNN	0.606	0.367	0.353	0.477	0.697	0.007
	LGB	0.541	0.292	0.335	0.584	0.765	1.953
	NNFastAi	0.527	0.277	0.332	0.605	0.782	4.748
	NNTorch	0.517	0.267	0.316	0.619	0.788	5.239
	RF	0.523	0.274	0.317	0.610	0.783	0.521
	XGB	0.542	0.294	0.332	0.581	0.766	2.103
	XT	0.510	0.260	0.325	0.630	0.795	0.356

Table 7. Simulation results incorporating feature selection approaches into the developed model.

of 30.372 and 0.356 seconds, respectively. However, model XT emerged as demonstrably superior in terms of training efficiency.

Feature selection analysis

A computational experiment was conducted to evaluate the effectiveness of different feature selection methods for TOC prediction using a dataset containing well-log measurements and corresponding TOC values. Feature selection can improve model performance and reduce training time in some cases. However, this task is challenging and there are many approaches in the literature. There are methods based on statistical tests, which use decision trees, among others. Three feature selection methods were compared: Boruta, Mutual Information (MI), and Recursive Feature Elimination (RFE). A total of 200 independent runs were performed for each feature selection method. In each realization, the data was shuffled, and each method was applied to select the most relevant variables. The dataset was then randomly divided into training and testing sets. The model was trained on the training set, and the testing set was used to calculate a performance metric. The MSE metric, described in Table 4, was used to compare the performance of the feature selection models.

Due to the randomization employed, each independent run yielded distinct results. A count indicator was used to track the variables selected in each run to identify the most relevant variables. If a variable was selected, the count returned 1; otherwise, it returned 0. If a variable was selected in all 200 runs, the count returned 200. Conversely, if a variable was selected in n runs, the count returned n. To simplify the decision-making process, the count was transformed into a percentage value, as presented in Table 8.

The variables obtained by the models differed due to their feature selection strategies. Table 2 shows that, of the seven variables, the most relevant for the Boruta method were acoustic time difference (AC), deep resistivity (RD), and uranium (U), thorium (Th), and potassium (K), appearing in 100% of the procedure's runs. The input variables gamma radiation (GR) and shallow resistivity (RS) were listed 18.5% and 18% of the time, respectively. In this context, for the Boruta method, the GR and RS variables were disregarded, resulting in a set with five variables (AC, K, RD, TH, and U).

The results for Mutual Information (MI) yielded three relevant variables (AC, GR, RS), while the RD variable was discarded as it appeared in only 29% of the runs. The remaining variables were not selected by the strategy. Interestingly, MI excludes features like RD, TH, and U, which Boruta considered important. This difference highlights the varying approaches of these methods. The Recursive Feature Elimination (RFE) process yielded the variables AC, K, and U. The RD variable was disregarded as it appeared in only 9.5% of the independent runs.

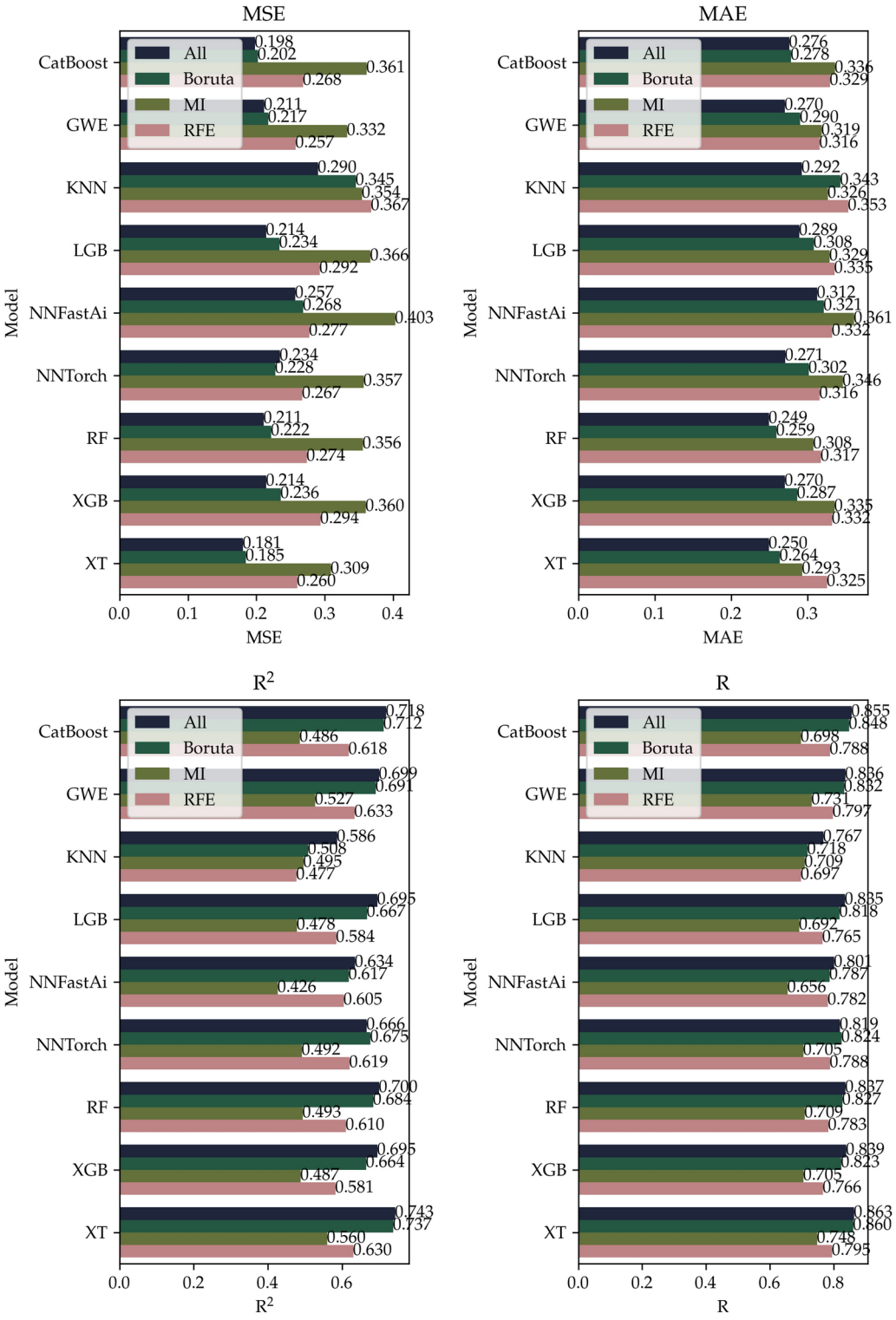

Fig. 7. Comparison of performance metrics for reduced datasets. The original dataset has seven input features as described in Table 1, the Boruta feature selection yielded 5 features, and MI and RFE have 3 features each.

Figure 8 shows a boxplot of the comparative performance measures using the reduced datasets with an XGB baseline model. The results indicate that the dataset produced by the Boruta method yielded the best performance, which is expected since the set has more variables, allowing the model to work with more formations and thus generate a model with greater accuracy and predictability. The proposed computational experiment shows that the Boruta method consistently outperformed the other methods, reducing the model complexity and the performance of ML models in this domain. However, it is important to note that the number of features selected

Feature selection	AC	K	RD	TH	U	GR	RS
Boruta	100.0	100.0	100.0	100.0	100.0	18.5	18.0
MI	100.0	–	29.0	–	–	100.0	71.0
RFE	100.0	90.5	9.5	–	100.0	–	–

Table 8. Percentage of the number of iterations the features are selected.

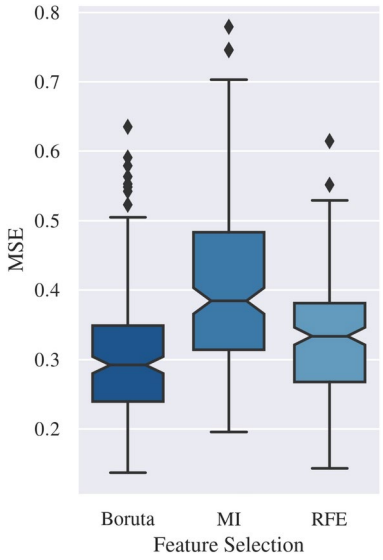

Fig. 8. Comparative performance of features selection methods using the reduced datasets.

by MI and RFE might be predefined or determined by the algorithm's stopping criteria, which is limited to 3. In addition, for MI and RFE, the user-defined parameters can impact the number of features selected. Following are some facts that may justify Boruta having a better result: Boruta selects strong and weak relevant features by comparing feature importance with random features, ensuring robustness to irrelevant and redundant features. It works well with high/low-dimensional datasets and correlated features, which can hinder MI and RFE. It makes use of statistical tests to provide more reliable feature selection and can handle non-linear relationships more effectively. RFE and MI often miss weak but relevant features or fail to capture all information, while Boruta's comprehensive approach tends to retain the most informative subset.

Each feature selection method used resulted in a different selection, some selected more, others less. It is important to evaluate the impact on the prediction as performed in the Subsection 3.2, in addition to using domain knowledge whenever possible[89,90], in this case the relationship between TOC and petrophysical features.

Feature importance analysis

Calculating the importance of features through permutation is a method verification tool that can be applied to supervised fitted models[71,91]. Using this approach makes nonlinear models convenient. Feature importance is established by assessing the decrease in a model's score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target value, so the decrease in the model score is an indication of how much the model depends on the feature. For example, a feature importance of 0.02 indicates that the predictive performance decreased by 0.02 when the feature was randomly shuffled. The higher the score for a feature is, the more critical it is to model performance.

Table 9 shows the coefficients calculated by analyzing the importance of the variables. The same coefficients are displayed on a bar graph in Fig. 9. Observing Table 9 and Fig. 9, it is possible to extract crucial information about the importance of each feature in the model. A detailed analysis of feature importance seeks to interpret the crucial role of each input variable in the context of AutoML modeling, aiming to improve the precision of forecasting the total organic carbon (TOC) content in oil wells.

The characteristic uranium (U) takes center stage, which is important at 0.7393. This eminence is justified by the strong correlation between TOC content and uranium content resulting from the incorporation of this element during the deposition of organic matter. The role of U is so remarkable that one can anticipate a considerable influence on TOC predictions, revealing the distinctiveness of diagenesis. Its impact manifests not only as a direct indicator of TOC but also as a marker of specific geological environments that may harbor higher concentrations of hydrocarbons.

The deep resistivity (RD) importance is 0.2986. This robust relationship between actual density and TOC content provides a solid basis for the contribution of RD to the predictions. This characteristic reflects variations

	Importance	std	*p*-value
U	0.751	0.087	$\leq 10^{-6}$
RD	0.233	0.032	$\leq 10^{-6}$
K	0.215	0.043	$\leq 10^{-6}$
RS	0.184	0.041	$\leq 10^{-6}$
AC	0.095	0.018	$\leq 10^{-6}$
GR	0.093	0.015	$\leq 10^{-6}$
TH	0.033	0.008	$\leq 10^{-6}$

Table 9. Importance of characteristics in the TOC prediction process.

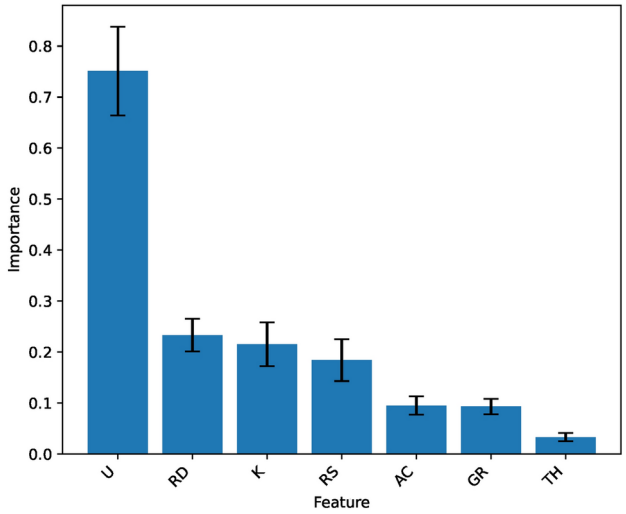

Fig. 9. Bar chart of feature importance. The black line in the blue bars represents the standard deviation of the calculated importance, and the height of the bars represents the mean.

in wellbore properties, which, in turn, reflect the presence and distribution of organic matter. The extent to which the actual density varies may indicate the presence of zones with higher or lower organic matter content, directly impacting the TOC estimates.

The gamma ray (GR) plays a notable role, with an importance of 0.1468. This characteristic is important because of its ability to reflect not only natural gamma ray but also radioactive elements such as potassium, thorium, and uranium. These elements are often associated with organic matter and sediment minerals and thus have complex relationships with TOC levels.

The acoustic (AC) characteristic is highlighted with an importance of 0.1319. A direct connection between electrical conductivity and TOC content indicates its influence. The unique feature of electrical conductivity is its ability to provide information about the electrical properties of sediment at different depths. This makes it possible to identify zones where organic matter or conductive minerals are more significant. The link between shear wave velocity and TOC content provides insight into the relationships between wellbore mechanical properties and TOC.

The shallow resistivity (RS) has significant numerical importance, taking the value of 0.1115. The shear wave velocity (Rs) plays a key role in predicting TOC levels in oil wells, contributing to the understanding of the mechanical and structural properties of sediment. Its importance stems from the intricate relationship between shear wave velocity and the physical and geological properties of the well. The relationship between shear wave velocity and TOC content is associated with the porosity distribution and the presence of organic matter. In some cases, organic matter can fill the spaces between sediment grains, altering the propagation of shear waves. Therefore, variations in shear wave velocity may be indicative of changes in organic matter distribution and porosity. The shear wave velocity contributes to the direct estimation of TOC and provides information on wellbore architecture and variations in physical properties across depths.

The input variables thorium (Th) and potassium (K) exhibit similar importance, reinforcing their role in the analysis. TH reveals nuances with importance equal to 0.0507, indicating the presence of thorium in shale rich in organic matter. Similarly, potassium, with an importance of 0.04741, reveals the influence of potassium on well composition. Thorium and potassium traits are complementary in predicting TOC. The characteristic K (potassium) provides important information about the mineral composition of the sediment. Potassium is present in several minerals, and its concentration can vary depending on the type of rock and geological conditions. Although its contribution is relatively moderate, the presence of potassium in association with other

radioactive elements, such as uranium and thorium, can influence the electrical properties and composition of the sediment. As a result, potassium (K) adds information about the complexity of the sedimentary environment. On the other hand, even though thorium has less prominent importance than other characteristics, it contributes to the understanding of the geological history of wells. Thorium is often associated with minerals that occur in sediments, and its presence may indicate certain sedimentary environments that favor the accumulation of organic matter. Therefore, including TH in the TOC prediction model allows us to capture nuances associated with specific geological contexts.

Discussion

This paper proposes an approach based on automated machine learning combined with feature selection approaches to investigate the prediction of total organic carbon (TOC) content. Considering the growing complexity of energy sources, the challenges in geochemistry and geophysics, and the exploration of energy resources in the energy transition era, this paper explored the potential of automated approaches to select the most suitable models and adjust their parameters to improve the accuracy of TOC predictions.

The proposed computational framework aimed to mitigate the limitations and bias associated with manual model selection and parameter setting approaches. The model consistently shows the potential to achieve reliable and consistent results, considering the complex interactions between models and underlying parameters. The automation of the TOC prediction process allowed the exploration of the search space, identifying the most promising combinations and, consequently, maximizing the predictive performance.

This approach is relevant because of the dynamic nature of the fields of geochemistry, geophysics and resource exploration[92–94]. With technological advances and the increasing availability of data, implementing automated methods has become essential for optimizing informed decision-making. Furthermore, the approach contributes to biases and subjectivities by reducing human intervention in the selection and configuration phases.

Recent studies that have developed ML models for various versatile applications have reported that hyperparameter tuning is crucial for ensuring proper model performance[95]. An effective fit can exploit the capabilities of simple models, resulting in competitive results. On the other hand, an inadequate fit can lead to a decline in the accuracy and robustness of models. An alternative to model tuning is the use of approaches that involve combining metaheuristics with ML models, resulting in hybrid models in which ML models benefit from automated search capabilities.

As the dataset size is limited, the application of complex models such as ensemble approaches increases the risk of overfitting, as the flexibility of these models makes them susceptible to capturing noise or irrelevant patterns in the data. Cross-validation helps estimate model performance (and reduces bias), but it may or may not prevent overfitting, especially if the data is limited or strongly unbalanced. In addition, early stopping, pruning, or limiting the number of ensemble trees can also further help reduce this risk by avoiding too many fittings. Also has the potential to enhance the dataset via data augmentation or synthetic data generation. The problem related to parameter search is formalized as an optimization problem in which grid search seeks to minimize metrics of interest, such as MSE (Mean Squared Error), while seeking the best combinations of hyperparameters[96,97]. In some scenarios, variable selection can be easily built into models. In these cases, in addition to coding the hyperparameters, the solutions are designed to include binary arrays that turn on or off the variables that feed the ML model[48,98]. More complex cases may involve sets of models, where the solution may indicate more than one learning model in a pipeline or linear combination strategy[99]. Other studies, such as[100,101], show the need for the development of integrated artificial intelligence systems in the area of petroleum engineering and propose the use of techniques such as Principal Component Analysis to reduce the dimensionality of the data improving the performance and reduce the processing time of the artificial neural network used to predict porosity and permeability. This final approach can generate accurate but overly complex models. In these contexts, the goal is to maximize the performance of the model by increasing its precision while trying to simplify the models.

These objectives may conflict, leading to the formulation of a multicriteria problem, that is, a problem that encompasses multiple[102] objectives. The solutions of interest, in this case, form a set of nondominated solutions known as the Pareto front. This front allows the decision maker to choose between more precise or simpler solutions, all equivalent to each other.

AutoML brings efficiency by automating model selection, tuning of hyperparameters, and feature engineering, but it also tends to be computationally intensive. Algorithms may vary in complexity and resource consumption based on their underlying functionality. As data sizes increase, consuming far more resources to train them, when training multiple models with a suite of algorithms, the process can be resource-intensive, often requiring CPU/GPU power and memory. In real-world situations, especially in the field, computational budgets are often limited, which can hinder detection. Running AutoML frameworks in those situations might require utilizing high-performance computers, cloud-based solutions, or optimized architectures, which can be avoided. Furthermore, some field applications require real-time predictions, making AutoML workflows inapplicable due to their generally slow work speed compared to the required decision time. The right use of this can be given for available future scenarios, which help in the later decisions. Therefore, a balanced approach is necessary, where the trade-offs between the power of AutoML and the practical constraints of the deployment environment are carefully considered.

The generalization of AutoML models to different geospatial and environmental conditions is a significant challenge. Although promising, AutoML frameworks usually need large, diverse datasets to train models and deliver decent performance on specific problems. In geoscientific use cases, we note that such data may be scarce, noisy, and specific to a region, resulting in models produced through AutoML not effectively generalizing to new, unknown environments without retraining or modification. In geoscience, professional expert knowledge is often lost in AutoML methods. Although AutoML may facilitate the efficiency of the model building process, it can simplify problems of a scientific nature by applying data-driven techniques without encompassing

relevant domain knowledge. Hence, although AutoML is a very promising opportunity, it is crucial to bear these limitations in mind when applying it to geoscientific problems to ensure the results remain scientific and applicable to the real world problems.

As the problems faced in areas such as geochemistry and energy resource exploration become more complex, the interpretability of ML models has gained prominence. Interpretability is critical for understanding and trusting model results, especially in domains where decision-making is critical, such as the oil and gas industry. Models that predict total organic carbon (TOC) content in reservoirs can be black boxes, and their decisions can be difficult to understand and explain, making it difficult for domain experts to trust and practically adopt predictions. In these scenarios, interpretable models gain relevance, which can explain how a certain prediction was reached. One of the techniques to improve the interpretability of AutoML models is the analysis of the importance of variables, as developed in this study, which provides insights into which variables have the greatest influence on model decisions. Understanding and explaining model decisions is critical for ensuring expert confidence and the practical utility of predictions, and understanding and explaining model decisions is a key aspect of research and model development in these domains. However, in some cases, interpretable models can be oversimplified to favor interpretability, leading to a loss of performance in terms of predictive ability. Therefore, a balanced approach that considers both performance and interpretability is needed to address complex problems.

AutoML models can help experts predict total organic carbon (TOC) content, as they allow adaptation to different geological contexts. They can be trained on a variety of data from different geological contexts, allowing experts to use the learned knowledge in a wide range of scenarios. Determining TOC in core rocks can be a time-consuming process involving several steps. Interpreting these results is critical for understanding the hydrocarbon-generating potential of rock formations. Using ML models to support geologists results in greater agility in the analysis process because, as new data are collected, the ML models can be updated and refined easily. This allows experts to track changes in geological conditions and continually improve forecasts.

Conclusion

This study evaluated the use of three selection features techniques combined with nine AutoML models for a TOC modeling problem with data collected in five shale formations in different geological basins: Yanchang shale in the Ordos Basin, the Shahejie shale in the Bohai Bay Basin, Longmaxi shale in the Sichuan Basin, the Shanxi and Taiyuan shales in the North China Basin. The AutoML approach allows preprocessing, grid searches for method hyperparameters, and evaluation in an integrated framework. The performance metrics RMSE, MSE, MAE, R^2 and R were used to assess the errors in the model's predictions.

The main conclusions of the research are as follows:

- The best result was obtained by AutoML-generated XT model with a correlation coefficient of 0.863 and a mean squared error (MSE) of 0.1806 for TOC prediction.
- The Greedy Weighted Ensemble model generated by the others AutoML models was evaluated and demonstrated a good performance but it took longer to train compared to the other models.
- The integration of feature selection within the AutoML framework, particularly the Boruta method, represents a significant step towards developing more robust and interpretable data-driven models for complex geoscience problems.
- The proposed strategy provides a powerful data-driven method for real-world wellbore applications, assisting in data analysis and subsequent decision-making.

Although limited, the study presented in this paper represents a contribution to petrophysical and energy resource exploration. It provides an advanced methodological approach that can enrich research practices and informed decision-making in these dynamic fields. The findings presented here can facilitate a general discussion about the application of automated techniques in the oil and gas industry.

Data availability

The present investigation is based on a dataset available at https://doi.org/10.3390/en16104159. The source code is available to interested parties upon request from the corresponding author.

Received: 22 September 2024; Accepted: 19 February 2025
Published online: 27 March 2025

References

1. Hu, T. et al. Movable oil content evaluation of lacustrine organic-rich shales: Methods and a novel quantitative evaluation model. *Earth Sci. Rev.* **214**, 103545. https://doi.org/10.1016/j.earscirev.2021.103545 (2021).
2. Xi, Z., Tang, S., Zhang, S., Lash, G. G. & Ye, Y. Controls of marine shale gas accumulation in the eastern periphery of the Sichuan basin, South China. *Int. J. Coal Geol.* **251**, 103939. https://doi.org/10.1016/j.coal.2022.103939 (2022).
3. Wang, H. et al. Unsupervised contrastive learning for few-shot toc prediction and application. *Int. J. Coal Geol.* **259**, 104046. https://doi.org/10.1016/j.coal.2022.104046 (2022).
4. Tabatabaei, S. M. E., Kadkhodaie-Ilkhchi, A., Hosseini, Z. & Asghari Moghaddam, A. A hybrid stochastic-gradient optimization to estimating total organic carbon from petrophysical data: A case study from the Ahwaz oilfield, SW Iran. *J. Petrol. Sci. Eng.* **127**, 35–43. https://doi.org/10.1016/j.petrol.2015.01.028 (2015).
5. Sun, J. et al. Prediction of toc content in organic-rich shale using machine learning algorithms: Comparative study of random forest, support vector machine, and XGBoost. *Energies* **16**, 4159. https://doi.org/10.3390/en16104159 (2023).
6. Alqahtani, A. & Tutuncu, A. Quantification of total organic carbon content in shale source rocks: An eagle ford case study. In *SPE/AAPG/SEG unconventional resources technology conference*, URTEC–1921783 (URTEC, 2014).

7. Zhao, P., Mao, Z., Huang, Z. & Zhang, C. A new method for estimating total organic carbon content from well logs. *AAPG Bull.* **100**, 1311–1327 (2016).
8. Al-Saddique, M., Hamada, G. & Al-Awad, M. N. State of the art: Review of coring and core analysis technology. *J. King Saud Univ.-Eng. Sci.* **12**, 117–137 (2000).
9. McPhee, C., Reed, J. & Zubizarreta, I. *Core analysis: A best practice guide* (Elsevier, 2015).
10. Holdaway, K. R. & Irving, D. H. *Enhance oil and gas exploration with data-driven geophysical and petrophysical models* (John Wiley & Sons, 2017).
11. Tariq, Z. *et al.* A systematic review of data science and machine learning applications to the oil and gas industry. *J. Pet. Explor. Prod. Technol.* 1–36 (2021).
12. Wayo, D. D. K., Irawan, S., Satyanaga, A. & Kim, J. Data-driven fracture morphology prognosis from high pressured modified proppants based on stochastic-ADAM-RMSprop optimizers; tf.NNR study. *Big Data Cogn. Comput.* **7**, 57 (2023).
13. Kizayev, T. et al. Factors affecting drilling incidents: Prediction of suck pipe by XGBoost model. *Energy Rep.* **9**, 270–279 (2023).
14. Holdaway, K. R. *Harness oil and gas big data with analytics: Optimize exploration and production with data-driven models* (John Wiley & Sons, 2014).
15. Balaji, K. *et al.* Status of data-driven methods and their applications in oil and gas industry. In *SPE Europe featured at 80th EAGE conference and exhibition* (OnePetro, 2018).
16. Shukla, A. & Karki, H. Application of robotics in offshore oil and gas industry-a review part II. *Robot. Auton. Syst.* **75**, 508–524 (2016).
17. Gooneratne, C. P. et al. Drilling in the fourth industrial revolution-vision and challenges. *IEEE Eng. Manag. Rev.* **48**, 144–159 (2020).
18. Basu, S., Jones, A. & Mahzari, P. Best practices for shale core handling: Transportation, sampling and storage for conduction of analyses. *J. Mar. Sci. Eng.* **8**, 136 (2020).
19. Cook, J., Growcock, F., Guo, Q., Hodder, M. & van Oort, E. Stabilizing the wellbore to prevent lost circulation. *Oilfield Rev.* **23**, 26–35 (2011).
20. Feng, Y. & Gray, K. Review of fundamental studies on lost circulation and wellbore strengthening. *J. Petrol. Sci. Eng.* **152**, 511–522 (2017).
21. Feng, Y. & Gray, K. *Lost circulation and wellbore strengthening* (Springer, 2018).
22. Yuan, B. & Wood, D. A. A comprehensive review of formation damage during enhanced oil recovery. *J. Petrol. Sci. Eng.* **167**, 287–299 (2018).
23. Dusseault, M. B., Jackson, R. E. & Macdonald, D. *Towards a road map for mitigating the rates and occurrences of long-term wellbore leakage* (University of Waterloo, 2014).
24. Muehlenbachs, L. A dynamic model of cleanup: Estimating sunk costs in oil and gas production. *Int. Econ. Rev.* **56**, 155–185 (2015).
25. Mirzaei-Paiaman, A., Asadolahpour, S. R., Saboorian-Jooybari, H., Chen, Z. & Ostadhassan, M. A new framework for selection of representative samples for special core analysis. *Pet. Res.* **5**, 210–226 (2020).
26. Carvajal-Ortiz, H. & Gentzis, T. Critical considerations when assessing hydrocarbon plays using rock-eval pyrolysis and organic petrology data: Data quality revisited. *Int. J. Coal Geol.* **152**, 113–122 (2015).
27. Gomaa, S., Mongy, M., Emara, R., Fahmy, A. & Attia, A. Evaluating medium decision tree model, support vector machine rational quadratic gaussian process regression to estimate the total organic carbon of shale gas reservoirs. *Pet. Coal***66** (2024).
28. Bione, F. R. et al. Estimating total organic carbon of potential source rocks in the Espírito Santo Basin, SE Brazil, using XGBoost. *Mar. Pet. Geol.* **162**, 106765 (2024).
29. Silva, R. O., Saporetti, C. M., Yaseen, Z. M., Pereira, E. & Goliatt, L. An approach for total organic carbon prediction using convolutional neural networks optimized by differential evolution. *Neural Comput. Appl.* **35**, 20803–20817 (2023).
30. Mahmoud, A. A., Elkatatny, S., Ali, A. Z., Abouelresh, M. & Abdulraheem, A. Evaluation of the total organic carbon (TOC) using different artificial intelligence techniques. *Sustainability* **11**, 5643 (2019).
31. Tariq, Z., Mahmoud, M., Abouelresh, M. & Abdulraheem, A. Data-driven approaches to predict thermal maturity indices of organic matter using artificial neural networks. *ACS Omega* **5**, 26169–26181 (2020).
32. Schmoker, J. Determination of organic content of Appalachian Devonian shales from formation-density logs. *Am. Assoc. Pet. Geol. Bull.* **63**, 1504–1509 (1979).
33. Passey, Q., Creaney, S., Kulla, J., Moretti, F. & Stroud, J. A practical model for organic richness from porosity and resistivity logs. *Am. Asso. Petrol. Geol. Bull.* **74**, 1777–1794 (1990).
34. Zheng, D., Wu, S. & Hou, M. Fully connected deep network: An improved method to predict toc of shale reservoirs from well logs. *Mar. Pet. Geol.* **132**, 105205 (2021).
35. Chan, S. A. et al. Total organic carbon (TOC) quantification using artificial neural networks: Improved prediction by leveraging XRF data. *J. Petrol. Sci. Eng.* **208**, 109302 (2022).
36. Asante-Okyere, S., Marfo, S. A. & Ziggah, Y. Y. Estimating total organic carbon (TOC) of shale rocks from their mineral composition using stacking generalization approach of machine learning. *Upstream Oil Gas Technol.* **11**, 100089 (2023).
37. Zhang, H., Wu, W. & Wu, H. Toc prediction using a gradient boosting decision tree method: A case study of shale reservoirs in Qinshui Basin. *Geoenergy Sci. Eng.* **221**, 111271 (2023).
38. Zhu, L., Zhou, X., Liu, W. & Kong, Z. Total organic carbon content logging prediction based on machine learning: A brief review. *Energy Geosci.* **4**, 100098 (2023).
39. Li, Y., Zhang, Y., Li, Y., Zhang, Y. & Li, Y. Novel intelligent system for predicting and guiding biogas performance in industrial-scale garage dry fermentation. *Energy Fuels* **35**, 8255–8266. https://doi.org/10.1021/acs.energyfuels.1c01922 (2021).
40. Azadivash, A., Soleymani, H., Kadkhodaie, A., Yahyaee, F. & Rabbani, A. R. Petrophysical log-driven kerogen typing: Unveiling the potential of hybrid machine learning. *J. Pet. Explor. Prod. Technol.* 1–29, https://doi.org/10.1007/s13202-023-01688-1 (2023).
41. Jia, W., Zong, Z., Qin, D. & Lan, T. A method for predicting the toc in source rocks using a machine learning-based joint analysis of seismic multi-attributes. *J. Appl. Geophys.* **216**, 105143. https://doi.org/10.1016/j.jappgeo.2023.105143 (2023).
42. Wood, D. A. Predicting total organic carbon from few well logs aided by well-log attributes. *Petroleum* **9**, 166–182. https://doi.org/10.1016/j.petlm.2022.10.004 (2023).
43. Tewari, S. & Dwivedi, U. A novel automatic detection and diagnosis module for quantitative lithofacies modeling. In *Abu Dhabi international petroleum exhibition and conference*, D012S122R001 (SPE, 2018).
44. Tewari, S. & Dwivedi, U. Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. *Comput. Ind. Eng.* **128**, 937–947 (2019).
45. Tewari, S. & Dwivedi, U. A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies. *J. Pet. Explor. Prod. Technol.* **10**, 1849–1868 (2020).
46. Kaleem, W., Tewari, S., Fogat, M. & Martyushev, D. A. A hybrid machine learning approach based study of production forecasting and factors influencing the multiphase flow through surface chokes. *Petroleum* **10**, 354–371 (2024).
47. Tewari, S. & Dwivedi, U. A real-world investigation of TwinSVM for the classification of petroleum drilling data. In *2019 IEEE region 10 symposium (TENSYMP)*, 90–95 (IEEE, 2019).
48. Goliatt, L., Mohammad, R. S., Abba, S. I. & Yaseen, Z. M. Development of hybrid computational data-intelligence model for flowing bottom-hole pressure of oil wells: New strategy for oil reservoir management and monitoring. *Fuel* **350**, 128623. https://doi.org/10.1016/j.fuel.2023.128623 (2023).

49. Chen, H., Wang, T., Zhang, Y., Bai, Y. & Chen, X. Dynamically weighted ensemble of geoscientific models via automated machine-learning-based classification. *Geosci. Model Dev.* **16**, 5685–5701 (2023).
50. Mubarak, Y. & Koeshidayatullah, A. Hierarchical automated machine learning (AutoML) for advanced unconventional reservoir characterization. *Sci. Rep.* **13**, 13812 (2023).
51. Davy, N. et al. Leveraging automated deep learning (AutoDL) in geosciences. *Comput. Geosci.* **188**, 105600 (2024).
52. He, X., Zhao, K. & Chu, X. AutoML: A survey of the state-of-the-art. *Knowl.-Based Syst.* **212**, 106622 (2021).
53. Goliatt, L., Saporetti, C. & Pereira, E. Super learner approach to predict total organic carbon using stacking machine learning models based on well logs. *Fuel* **353**, 128682 (2023).
54. Rahaman, M. et al. Evaluation of tree-based ensemble learning algorithms to estimate total organic carbon from wireline logs. *Int. J. Innov. Comput. Inf. Control* **17**, 807–829 (2021).
55. Liu, X., Tian, Z. & Chen, C. Total organic carbon content prediction in lacustrine shale using extreme gradient boosting machine learning based on Bayesian optimization. *Geofluids* **2021**, 1–18 (2021).
56. Safaei-Farouji, M. & Kadkhodaie, A. Application of ensemble machine learning methods for kerogen type estimation from petrophysical well logs. *J. Petrol. Sci. Eng.* **208**, 109455 (2022).
57. Saporetti, C., Fonseca, D., Oliveira, L., Pereira, E. & Goliatt, L. Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. *Mar. Pet. Geol.* **143**, 105783 (2022).
58. Zhu, L. et al. Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves. *J. Geophys. Eng.* **15**, 1050–1061 (2018).
59. Rui, J., Zhang, H., Zhang, D., Han, F. & Guo, Q. Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. *J. Petrol. Sci. Eng.* **180**, 699–706 (2019).
60. Topsakal, O. et al. Utilization of machine learning for the objective assessment of rhinoplasty outcomes. *IEEE Access* **11**, 42135–42145. https://doi.org/10.1109/ACCESS.2023.3270438 (2023).
61. Vakhrushev, A. *et al.* Lightautoml: AutoML solution for a large financial services ecosystem. arXiv:2109.01528 (2021).
62. Zhang, H. et al. Artificial intelligence for the diagnosis of clinically significant prostate cancer based on multimodal data: A multicenter study. *BMC Med.* **21**, 270. https://doi.org/10.1186/s12916-023-02964-x (2023).
63. Palacios Salinas, N. R., Baratchi, M., van Rijn, J. N. & Vollrath, A. Automated machine learning for satellite data: integrating remote sensing pre-trained models into Automl systems. In *Joint European conference on machine learning and knowledge discovery in databases*, 447–462 (Springer, 2021).
64. Chauhan, K. *et al.* Automated machine learning: The new wave of machine learning. In *2020 2nd international conference on innovative mechanisms for industry applications (ICIMIA)*, 205–212 (2020).
65. Alsharef, A., Aggarwal, K., Sonia, A. A. A., Kumar, M. & Mishra, A. Review of ml and AutoML solutions to forecast time-series data. *Arch. Comput. Methods Eng.* **29**, 5297–5311. https://doi.org/10.1007/s11831-022-09765-0 (2022).
66. Ke, G. *et al.* LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** (2017).
67. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **31** (2018).
68. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
69. Chen, T. *et al.* XGBoost: Extreme gradient boosting. *R package version 0.4-2* **1**, 1–4 (2015).
70. Bentley, J. L. Multidimensional binary search trees used for associative searching. *Commun. ACM* **18**, 509–517 (1975).
71. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
72. Erickson, N. *et al.* AutoGluon-Tabular: Robust and accurate AutoML for structured data. arXiv:2003.06505 (2020).
73. Yehia, T., Wahba, A., Mostafa, S. & Mahmoud, O. Suitability of different machine learning outlier detection algorithms to improve shale gas production data for effective decline curve analysis. *Energies* **15**, 8835 (2022).
74. Yehia, T., Wahba, A., Mostafa, S. & Mahmoud, O. Machine learning outlier detection algorithms for enhancing production data analysis of shale gas. *Fundam. Res. Appl. Phys. Sci.* **4**, 127–163 (2023).
75. Coelho, M. H., Bittencourt, O. O., Morelli, F. & Santos, R. Método para a classificação de áreas queimadas baseado em aprendizado de máquina automatizado. *Anais do Computer on the Beach* **13**, 029–036 (2022).
76. Kursa, M. B., Jankowski, A. & Rudnicki, W. R. Boruta-a system for feature selection. *Fund. Inform.* **101**, 271–285 (2010).
77. Stoppiglia, H., Dreyfus, G., Dubois, R. & Oussar, Y. Ranking a random feature for variable and feature selection. *J. Mach. Learn. Res.* **3**, 1399–1414 (2003).
78. Christ, M., Kempa-Liehr, A. W. & Feindt, M. Distributed and parallel time series feature extraction for industrial big data applications. arXiv:1610.07717 (2016).
79. Prasad, R., Deo, R. C., Li, Y. & Maraseni, T. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *CATENA* **177**, 149–166 (2019).
80. Ahmed, A. M. et al. Deep learning hybrid model with Boruta-random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *J. Hydrol.* **599**, 126350 (2021).
81. Lawal, I. et al. Application of Boruta algorithms as a robust methodology for performance evaluation of CMIP6 general circulation models for hydro-climatic studies. *Theoret. Appl. Climatol.* **153**, 113–135 (2023).
82. Hu, Z., Bao, Y., Xiong, T. & Chiong, R. Hybrid filter-wrapper feature selection for short-term load forecasting. *Eng. Appl. Artif. Intell.* **40**, 17–27 (2015).
83. Ren, K., Fang, W., Qu, J., Zhang, X. & Shi, X. Comparison of eight filter-based feature selection methods for monthly streamflow forecasting-three case studies on camels data sets. *J. Hydrol.* **586**, 124897 (2020).
84. Kraskov, A., Stögbauer, H. & Grassberger, P. Erratum: Estimating mutual information [phys. rev. e 69, 066138 (2004)]. *Phys. Rev. E* **83**, 019903 (2011).
85. Ross, B. C. Mutual information between discrete and continuous data sets. *PLoS ONE* **9**, e87357 (2014).
86. Yan, K. & Zhang, D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sens. Actuators B Chem.* **212**, 353–363 (2015).
87. Lian, W. et al. An intrusion detection method based on decision tree-recursive feature elimination in ensemble learning. *Math. Probl. Eng.* **2020**, 1–15 (2020).
88. Faysal, J. A. *et al.* XGB-RF: A hybrid machine learning approach for IoT intrusion detection. In *Telecom*, **3**, 52–69 (MDPI, 2022).
89. Yehia, T., Gasser, M., Ebaid, H., Meehan, N. & Okoroafor, E. R. Comparative analysis of machine learning techniques for predicting drilling rate of penetration (ROP) in geothermal wells: A case study of forge site. *Geothermics* **121**, 103028 (2024).
90. Yehia, T., Gasser, M., Ebaid, H., Meehan, N. & Okoroafor, E. R. A comparative analysis of machine learning techniques for geothermal wells' drilling rate of penetration (ROP) prediction. In *Unconventional resources technology conference, 17–19 June 2024*, 429–448 (Unconventional Resources Technology Conference (URTeC), 2024).
91. Saporetti, C., Fonseca, D., Oliveira, L., Pereira, E. & Goliatt, L. Machine learning with model selection to predict toc from mineralogical constituents: Case study in the Sichuan basin. *Int. J. Environ. Sci. Technol.* **20**, 1585–1596 (2023).
92. Klunk, M. A. et al. Geochemical modeling of diagenetic reactions in Snorre field reservoir sandstones: A comparative study of computer codes. *Br. J. Geol.* **45**, 29–40 (2015).
93. Zhu, G. et al. Formation mechanism and geochemical characteristics of shallow natural gas in heavy oil province, China. *Sci. China, Ser. D Earth Sci.* **51**, 96–106 (2008).
94. Magalhães, A. J. C. et al. Sequence stratigraphy of clastic and carbonate successions: Applications for exploration and production of natural resources. *Br. J. Geol.* **51**, e20210014 (2021).

95. Saporetti, C. M., da Fonseca, L. G. & Pereira, E. A lithology identification approach based on machine learning with evolutionary parameter tuning. *IEEE Geosci. Remote Sens. Lett.* **16**, 1819–1823. https://doi.org/10.1109/LGRS.2019.2911473 (2019).
96. Goliatt, L., Saporetti, C., Oliveira, L. & Pereira, E. Performance of evolutionary optimized machine learning for modeling total organic carbon in core samples of shale gas fields. *Petroleum* (2023).
97. Martinho, A. D., Hippert, H. S. & Goliatt, L. Short-term streamflow modeling using data-intelligence evolutionary machine learning models. *Sci. Rep.* **13**, 13824. https://doi.org/10.1038/s41598-023-41113-5 (2023).
98. Basílio, S. C. A., Saporetti, C. M., Yaseen, Z. M. & Goliatt, L. Global horizontal irradiance modeling from environmental inputs using machine learning with automatic model selection. *Environ. Dev.* **44**, 100766. https://doi.org/10.1016/j.envdev.2022.100766 (2022).
99. Boratto, T. H., Saporetti, C. M., Basilio, S. C., Cury, A. A. & Goliatt, L. Data-driven cymbal bronze alloy identification via evolutionary machine learning with automatic feature selection. *J. Intell. Manuf.* 1–17 (2022).
100. Mohaghegh, S. D. Recent developments in application of artificial intelligence in petroleum engineering. *J. Petrol. Technol.* **57**, 86–91 (2005).
101. Moghadasi, L., Ranaee, E., Inzoli, F. & Guadagnini, A. Petrophysical well log analysis through intelligent methods. In *SPE Norway subsurface conference*, D011S001R002 (SPE, 2017).
102. Gotardelo, D. & Goliatt, L. Multi-objective optimization of portfolio selection involving non-convex attributes in an anti-fragile perspective. *Evol. Syst.* 1–13 (2023).

Author contributions

Bruno da Silva Macêdo: data curation and software, validation, project administration, writing - review & editing. Dennis Delali Kwesi Wayo: methodology, validation and visualization, writing - original draft, review & editing. Deivid Campos: software, validation, writing - review & editing. Rodrigo Barbosa De Santis: data curation and software, methodology, validation. Alfeu Dias Martinho: validation, writing - review & editing. Zaher Mundher Yaseen: conceptualization, methodology, supervision, writing - review & editing. Camila M. Saporetti: methodology, validation, writing - review & editing. Leonardo Goliatt: conceptualization, methodology, supervision, writing - review & editing. All authors reviewed the manuscript.

Declarations

Competing interests

The author(s) declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.D.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.