RESEARCH PAPER

Tree-based machine learning can determine lithofacies properties of reservoir rocks − Camal oil field, Yemen

Abbas M. Al-khudafi [a], Ghareb M. Hamada [b,*], Abdelrigeeb Al-Gathe [c], Ibrahim A. Farea [a], Salem O. Baarimah [c]

[a] Oil and Gas Engineering Department, Emirates International University, Sanaa, Yemen
[b] Oil and Gas Engineering Department, Arab Academy for Science, Technology & Maritime Transport, P.O. Box 1019, Alexandria, Egypt
[c] Department of Petroleum Engineering, Hadhramaut University, Mukalla, Yemen

Abstract

This study aims to assess the effectiveness of several decision tree machine techniques for identifying formation lithology. A total of 20 966 log data points from four wells were used to create the study's data. Lithology is determined using seven log parameters. The seven log parameters are the density log, neutron log, sonic log, gamma ray log, deep lateral log, shallow lateral log, and resistivity log. Different decision tree-based algorithms for classification approaches were applied. Several typical machine learning models, namely the, Random Forest. Random trees, J48, reduced-error pruning decision trees, logistic model trees, and Hoeffding Tree were assessed using well-logging data for formation lithology prediction. The obtained results show that the random forest model, out of the proposed decision tree models, performed best at lithology identification, with precession, recall, and *F* score values of 0.913, 0.914, and 0.913, respectively. Random trees came next with average precision, recall, and *F*1 score of 0.837, 0.84, and 0.837, respectively, the J48 model came in third place. The Hoeffding Tree classification model, however, showed the worst performance. We conclude that boosting strategies enhance the performance of tree-based models. Evaluation of the prediction capability of models is also carried out using different datasets.

Keywords: Decision tree, Lithofacies prediction, Machine learning, Modeling, Well logging

1. Introduction

Lithology must be established using well-log data to explore and produce petroleum. The lithology model of a reservoir can be created by quantitative analysis of logging data. The high cost of drilling cores limits the amount of required logging data. Due to the intricacy of lithology, the distributions of logging data from distinct lithologies overlap, expanding the number of possible identifications. Thus, it is essential to use methods that provide an accurate means of forecasting lithology.

Researchers have recently become more interested in applying machine-learning approaches to forecast different types of lithology.[1–3] These approaches to lithology identification based on machine learning make an effort to train a multiclass classifier model based on a large amount of labeled well-logging data with logging curves, such as gamma ray, resistivity logs, sonic logs, neutron logs, and density logs.

Various machine-learning approaches have been proposed for the lithology classification problem. In lithological identification using logging data points, an artificial neural network (ANN) was first used to classify lithology.[4,5] Support vector machine (SVM) was utilized[6] to classify the lithology with logging data points and accurately identify the lithology facies of heterogeneous sandstone reservoirs.

Received 12 February 2024; revised 26 June 2024; accepted 8 July 2024.
Available online 16 September 2024

* Corresponding author. P.O.Box 1019.
E-mail address: ghareb.hamada@aast.edu (G.M. Hamada).

https://doi.org/10.62593/2090-2468.1041

Different types of multi-classification SVM were applied to identify volcanic lithology with well-log data.[7] Random Forest was utilized to predict lithological mapping based on geophysical and geochemical data.[8] In the field of spatial modeling and classification based on log data. The novel hybrid inferential system called ANN-hidden Markov models for lithofacies classification. Approaches to modeling the rock lithology were developed by using recurrent neural networks were used.[9,10] An ANN model to identify the lithology of a layer as it was being drilled using neighboring well data and real-time drilling data[11] from wells in the South Pars gas field. Using data from the Daniudui and Hangjinqi gas fields, five common machine learning techniques – Naïve Bayes, SVM, RF, ANN, and Gradient Tree Boosting – were assessed for detection of formation lithology.[12]

Conventional single classification algorithms such as decision trees, SVM, and Bayes were developed to determine the lithology of the Longqian region of China using well logs.[13] In order to predict the geological facies using well-log data in the Anadarko Basin, Kansas, supervised learning algorithms, unsupervised learning algorithms, and a neural network machine learning algorithm were presented.[14] Generative adversarial networks were presented to recreate thin section images and identify carbonate lithology.[15] An extreme gradient boosting and Bayesian optimization classifier were proposed for identifying the lithology of the Daniudui and Hangjinqi gas fields.[16] Three machine learning algorithms were presented to determine the lithology while drilling. Neural networks, RF, extreme gradient boosting tree algorithms, and one-versus-one SVMs) are used to create machine learning.[17] A coarse-to-fine architecture that incorporates outlier detection, multiclass classification, and a tree-based classifier is suggested for identifying the lithology using two actual well-logging data sets.[18] A novel hybrid framework combining ANNs and hidden Markov models for lithological sequence classification proposed by Feng.[19]. Coal pay zones were predicted using a variety of machine learning algorithms (LR, SVM, ANN, RF, and extreme gradient boosting tree) and data manipulation methods (NROS and SMOTE).[20] Bidirectional gated recurrent units and a conditional random field layer (Bi-GRU-CRF) are the models used in the lithological sequence classification technique that was proposed by Liu et al.[21] The performance of the gradient boosting decision tree model, which was validated in comparison with the ANN, SVM, AdaBoost, and RF classifiers, was demonstrated by Zou et al.[22] A Gray Wolf Optimization Algorithm-based automatic identification system for lithology logging has been presented.[23] A study by Abid et al.[24] examined the Khurmala Formation in Iraq, focusing on its paleoenvironment and microfacies composition. The study identified a diverse assemblage of microfacies, including coral-algal wackestone, foraminiferal-peloidal packstone, and grainstone. The presence of calcareous green algae suggests a potential link to sea-level changes. Martyushev et al.[25] explored carbonate reef reservoirs in the Upper Devonian oil fields of Russia, revealing a cyclic nature controlled by tectonic movements and sedimentation processes. The study also identified fracture types within the void space, including calcite-healed fractures that may affect fluid filtration. Makarian et al.[26] analyzed poroelastic media in a carbonate formation in southwest Iran, revealing that water saturation increases compressional wave velocity while decreasing shear wave velocity within the porous media.[27] examined the pore properties of various lithofacies within the Lianggaoshan Formation, a significant shale oil producer in Northeast Sichuan. ORLAS and OLLSS exhibited superior reservoir qualities compared to SM and FS, but siliceous minerals negatively impacted reservoir properties. Characterization of reservoirs and evaluating the efficacy of machine learning models in predicting lithofacies across diverse geological settings is paramount. This research investigates the application of various decision tree machine learning algorithms for the automated identification of rock formations (lithology) based on well-log data. The primary objective is to evaluate the effectiveness of these techniques in accurately classifying different lithological types.

1.1. Used well-logging data

The evaluation employed a dataset of 20 966 well data points, including log and cutting data from four wells in the Camal oil field. This dataset encompassed seven logging parameters [density log (RHOB), neutron log (NPHI), sonic log (DT), gamma ray log (GR), deep latero log (LLD), shallow latero log (LLS), and resistivity log (ML)] with corresponding depths. The output class to be identified is the type of lithology (shale, sand, sandstone, limestone, or dolomite). The range of the seven feature parameters is listed in Table 1. To assess the model's performance, we conducted evaluations on three distinct datasets, where each dataset reflected variations in the input parameters.

2. Machine learning models

To classify lithology in this study, we employed six machine learning algorithms based on decision

Table 1. Range of parameters for lithology classification.

Parameters	Maximum	Minimum	SD	Mean
ML	1952.27	0.23	273.34	112.99
LLD	2064.76	0.23	63.72	29.74
LLS	2064.76	0.22	100.03	33.60
Depth	6100	520	1555	3421
GR	139.37	7.87	21.36	43.69
RHOB	2.95	1.94	0.18	2.28
NBHI	0.45	0.01	0.10	0.27
DT	141.76	38.71	17.87	91.54

trees: Random Forest (RF), Random Trees (RT), J48, Reduced-Error Pruning Trees (REPT), Logistic Model Trees (LMT), and Hoeffding Trees (HT). Fig. 1 illustrates the proposed lithology classification methods.

2.1. Decision tree

Three nodes make up a decision tree, which is a classification method: the leaf node, the branch (edge or link), and the root node. The test conditions for various attributes are represented by the root, all possible test outcomes are represented by the branch, and the labels of the classes to which the leaf nodes belong are present. The beginning of the tree sometimes referred to as the top of the tree, is home to the root node. A decision tree is a hierarchical decision support model that uses a tree-like model of decisions and their potential repercussions, such as utility, resource costs, and chance event outcomes. It's one method of presenting an algorithm with just conditional control statements. In operations research, decision analysis, in particular, decision trees are frequently utilized.

2.2. Random forest

Known also as random decision forests, random forests are an ensemble learning technique that builds a large number of decision trees during the training phase for tasks like regression and classification. The class that the majority of the trees choose is the random forest's output for classification problems. The mean or average prediction

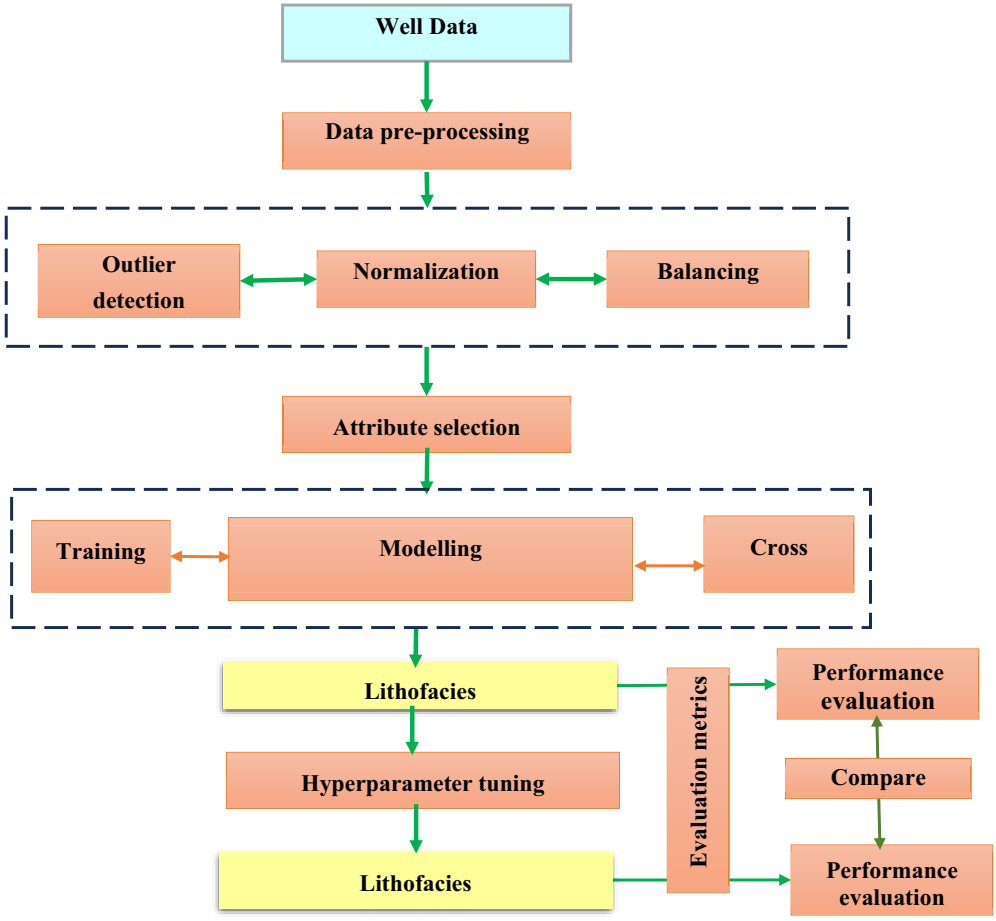

Fig. 1. Workflow for evaluation machine learning model.

made by each tree is returned for regression tasks.[28] The tendency of decision trees to overfit their training set is compensated for by random decision forests. Although they are less accurate than gradient-boosted trees, random forests still perform better than choice trees in most cases. The performance of the system may be contingent upon the characteristics of the data it processes.

2.3. Reduced-error pruning decision tree

In machine learning and search algorithms, pruning is a data compression approach that minimizes the size of decision trees by eliminating nonessential and redundant portions, for instance, classification (Matti, 2003). Pruning decreases overfitting, which lowers the complexity of the final classifier and increases predictive accuracy. Reduced error pruning is one of the most straightforward types of pruning. Every node, starting from the leaves, gets swapped out for its most popular class. The adjustment is retained if there is no impact on the prediction accuracy. Reduced mistake trimming gives performance and simplicity benefits, although being a little naïve.

2.4. Logistic model tree

Combining logistic regression (LR) and decision tree learning, the logistic model tree is a classification model that comes with a corresponding supervised training algorithm.[29] The concept of a logistic model tree is derived from the previous concept of a model tree, which is a decision tree with linear regression models at the leaves that generates a piecewise linear regression model instead of the piecewise constant model that would be produced by regular decision trees with constants at the leaves.[29]

2.5. Hoeffding Tree

One decision tree learning technique for classifying stream data is the Hoeffding Tree algorithm. An application of the incremental decision tree algorithm is the Hoeffding Tree. Originally, it was used to monitor clickstreams on the internet and build models to forecast which hosts and websites a user is most likely to visit. It usually produces a decision tree that is almost the same as that of standard batch learners and runs in sublinear time. It makes use of Hoeffding Trees, which make use of the fact that selecting the best splitting attribute is frequently possible with a small sample size. The Hoeffding bound, often known as the additive Chernoff bound, provides mathematical support for this theory.

2.6. J48 classifier

The c4.5 algorithm, developed by Ross Quinlan, is a prominent method for generating decision trees in machine learning. It falls under the category of information-theoretic classification algorithms, utilizing information gain to construct the tree structure. C4.5 builds upon Quinlan's earlier id3 algorithm, also known as j48. The primary function of c4.5 lies in classification tasks, where it leverages decision trees to assign data points to specific categories.[30]

3. Data preprocessing of well logs

Using seven logging features – density log, neutron log, sonic log, gamma ray log, deep latero log, shallow latero log, and resistivity log – a total of 20 966 well log data points were used for lithology classification.

3.1. Outlier removal

One of the main challenges is the presence of outliers and extremes in the dataset, which can deteriorate the performance of classifiers. Thus, the technique of unsupervised learning was applied to identify outliers within the dataset. These data samples might have come from contaminated or incorrectly entered logging parameters by hand. Finding data samples that differ from the distribution of the majority of data is the goal of outlier identification. Outliers and extremes sometimes deteriorate the performance of classifiers that cannot be used in a dataset. For this purpose, the interquartile range (IQR) and local outlier factor (LOF) filters were applied. The IQR filter detects outliers and extreme values. Then, the filter removal with value was implemented to remove outliers and extremes from data sets. The IQR filter is better than other available filters because it is a robust measure of variability that is not affected by extreme values or outliers. Additionally, it may be applied to a variety of datasets and is simple to use. LOF identifies an outlier based on the local neighborhood, which means it considers the density of the neighborhood to identify an outlier. Because it can detect outliers in a dataset that would not be outliers in another part of the dataset, the LOF algorithm outperforms alternative filters that are currently on the market. The LOF is shown to perform better for anomaly detection than many other methods and can also be utilized to construct a distinct dissimilarity function. Experiments were

conducted to evaluate the performance of both filters. According to the results, all classifiers have a higher prediction accuracy for the LOF filter.

3.2. Manage imbalanced dataset

Another challenge is the imbalanced dataset, which can lead to overfitting or underperformance of the model. To address imbalanced data and prevent overfitting or underperformance, we applied the Synthetic Minority Over-Sampling Technique (SMOTE). By increasing the proportion of minority instances in the dataset, this technique maintained balance and enhanced algorithm performance. We employed the SMOTE function[31] specifically to tackle class imbalance issues related to different lithology types, enhancing lithology prediction model performance. The application of the SMOTE method improved the model's performance. For the random forest model, for instance, oversampling raised accuracy from 88.2 to 92.1%.

3.3. Normalization

Data normalization is also a crucial step in the analysis, as logging indicators have varying dimensions. Since logging indicators have varying dimensions, we performed data normalization after data collection, mining, and quality control. This step ensures consistency and allows us to combine dimensionless data to create new analysis indicators. All of the dataset's numerical values were standardized to fall between 0 and 1 before the machine learning model was trained.

4. Predictive model building

Building the model for lithofacies prediction involved several steps, including data preprocessing, feature selection, model training, hyperparameter tuning, and validation. The dataset used consisted of well log and cutting data points from four wells, with seven logging parameters and corresponding depths. Data preprocessing involved outlier removal, normalization, and attribute selection. Outlier removal was done using unsupervised learning to identify and remove data samples that might have come from contaminated or incorrectly entered logging parameters. Normalization was performed to ensure consistency and allow for the combination of dimensionless data to create new analysis indicators. All numerical values in the dataset were standardized to fall between 0 and 1 before training the machine learning model. Feature selection was carried out to evaluate the importance of the features for the prediction models. Four algorithms were recommended in conjunction with rankers, which rank attributes by their evaluations. The major features contributing to lithology prediction were determined based on attribute rank. Different decision tree-based algorithms for classification approaches were applied. The prediction model was trained using the training dataset (80%), and it was tested using the test dataset (20%). The classification models were also constructed using a 10-fold cross-validation technique. Hyperparameters were tuned using a 10-fold cross-validation method to optimize machine-learning models for lithology identification. The best hyperparameter set for these models was determined by evaluating the influence of various hyperparameters on model performance. The optimal hyperparameters were used to construct the model classifier using 10-fold cross-validation, with nine subsets of the training datasets chosen for model training and hyperparameter tuning and one subset used for model validation. A cross-validation accuracy curve was utilized to find the best hyperparameters for the tree model. Validation techniques employed to ensure the robustness and generalization of the predictive models included 10-fold cross-validation and performance evaluation using various evaluation metrics. Every classification model was assessed using 10-fold cross-validation.

4.1. Hyperparameters

Hyperparameters are parameters that control the learning process in machine learning models. Unlike other parameters, such as node weights, which are learned during training, hyperparameters are set beforehand.[32] They can be categorized as model hyperparameters, which influence model selection, or algorithm hyperparameters, which affect the learning process's speed and quality. Model hyperparameters include factors like neural network topology and size, while algorithm hyperparameters encompass settings like learning rate, batch size, and mini-batch size. Different machine learning algorithms require specific hyperparameters, and tuning them is crucial for adapting models to specific datasets.[33] Tree depth and the total number of trees in a random forest are two instances of hyperparameters for tree models, and learning-related settings like the learning rate, batch size, and mini-batch size.

4.2. Tuning hyperparameter

This study utilized hyperparameter tuning to optimize machine-learning models for lithology identification.

A 10-fold cross-validation method was employed to find the best hyperparameter set for these models. This approach evaluated the influence of various hyperparameters on model performance and emphasized the importance of hyperparameter tuning in machine learning. The optimal hyperparameters for the tree model were determined based on the cross-validation results.

In order to construct the model classifier, 10-fold cross-validation was used, and the hyperparameters were optimized. A 10-fold cross-validation procedure was employed. The training data was randomly divided into 10 equally sized folds. In each iteration, nine folds were used to train and tune the model's hyperparameters, while the remaining fold was reserved for validation.

4.3. Attribute selection

Attribute selection is also an important preprocessing technique for quality control and data mining. The impact of parameter correlation on lithology detection model performance is investigated. We hypothesize that by employing data mining techniques for feature selection, we can identify the most influential parameters for accurate lithology classification. This approach will not only reduce the dimensionality of the input data, mitigating the risk of overfitting but also enhance the overall forecasting accuracy of the model. In this study, the importance of the features was evaluated for the prediction models. For selecting the log parameters, four algorithms were recommended in conjunction with rankers, which rank attributes by their individual evaluations. According to attribute rank, the major features contributing to lithology prediction were determined. The algorithms used include InfoGain, Relief, and OneR. The results are presented in Fig. 2.

As indicated in Fig. 2, analysis of feature importance scores suggests that depth is the most prominent factor influencing the model's predictions. This implies a prioritization of depth for accurate model outputs. However, it is crucial to acknowledge the relative nature of feature importance. Even features like NBHI, with the lowest score, might contribute to the model's decision-making process. Furthermore, limitations inherent to feature importance scores warrant consideration. These scores merely quantify a feature's internal significance to the model, not necessarily its real-world relevance. A seemingly important feature could be highly correlated with another that holds the true causal relationship with the outcome variable. To investigate the models' generalizability in predicting

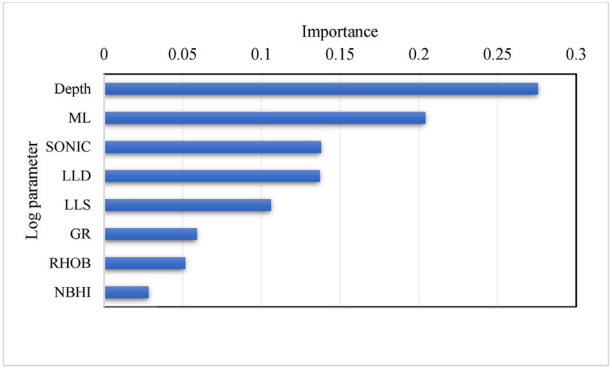

Fig. 2. Visualization of feature importance scores.

across diverse log parameter settings, we employed classification on eight functional datasets (Table 2).

Fig. 3 summarizes the interactions between various well-logging features and the various modeling performances. There are no statistically significant differences among the other algorithms in terms of their sensitivity to these properties, except for the HT model, which shows poor overall precision. Remarkably, for all combinations of input variables, RF model consistently produces optimal results across all sets. The relationship between different model accuracies and the quantity of log parameters is shown in Fig. 4. This figure explores the relationship between the number of logging parameters employed and the resulting accuracy of lithology prediction models.

Results in Fig. 4 indicate that a core set of four well logs, namely density, neutron, sonic, and gamma-ray, can be effectively utilized for lithology identification. The average accuracy across all models exhibited a positive correlation with the number of logging parameters, reaching a peak at six parameters. A slight decline in accuracy was observed with seven parameters. These findings highlight the importance of incorporating a sufficient number of logging parameters during model development. However, it is crucial to avoid overfitting by employing a balanced parameter selection strategy. The optimal number of parameters is likely dataset-

Table 2. Functional form characteristics of datasets.

Datasets	Log parameters
1	Depth, RHOB, GR, LLD
2	Depth, RHOB, GR, LLD, ML
3	Depth, LLD, ML
4	Depth, DT, LLD, LLS, ML
5	Depth, NBHI, RHOB, GR
6	Depth, NBHI, RHOB, GR, DT
7	Depth, RHOB, GR, LLS
8	Depth, RHOB, GR, LLS, ML

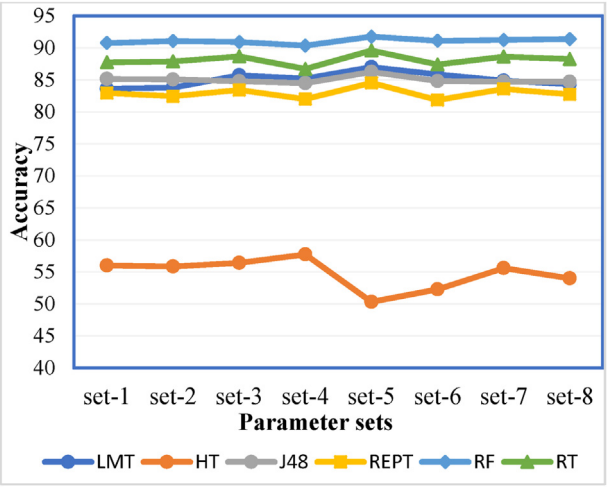

Fig. 3. Performance of model with different parameter sets.

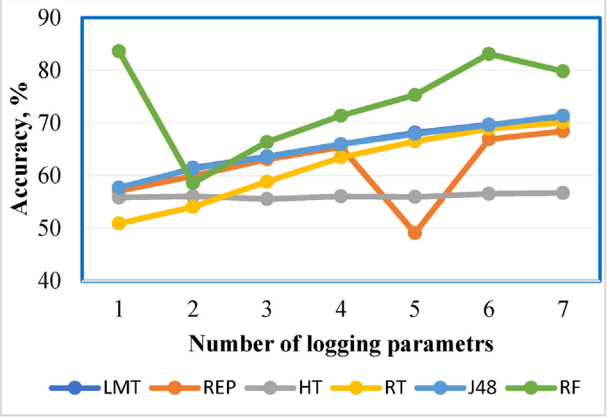

Fig. 4. Influence of logging parameters on lithology prediction.

dependent and model-dependent. Nevertheless, this study suggests that using approximately six logging parameters serves as a valuable starting point. The observed improvement in accuracy with an increasing number of parameters can be attributed to the enhanced information content provided by these parameters, enabling the models to make more precise predictions about the rock formations.

5. Results and discussion

5.1. An evaluation of model performance using metrics

In this study, various evaluation metrics were used to assess the performance of classification models. These metrics included classification accuracy (E), precision (Pr), recall (R), F-measure (F1), ROC area, and the PRC area in order to more thoroughly assess the effectiveness of the learning model and the impact of lithology identification. Every classification model was assessed using 10-fold cross-validation. Table 3 presents the performance metrics for the evaluated models.

Based on our investigation of classifier performance, we found the random forest model to be the standout performer (Table 3). This is evidenced by its exceptional classification accuracy, achieving a remarkable 92% during cross-validation and a very close 91.35% for training data. The RF model's strength lies in its ability to accurately identify sandstone and dolomite samples, as demonstrated in Table 4. Following closely behind the RF model was the J48 model, securing an average precision of 0.848, recall of 0.85, and F-measure of 0.848 (Table 3). The remaining models exhibited lower overall performance. The HT model yielded the least desirable results, while the RT model positioned itself as the second-best alternative to the RF model.

This analysis underscores the effectiveness of the RF model for rock type classification, particularly for identifying sandstone and dolomite. Future research could delve deeper into the importance of features within the RF model to gain a more nuanced understanding of its decision-making process.

The results presented in Table 4 demonstrate that the RF model achieved superior performance compared to other classifiers in distinguishing specific lithology types, such as shale (Sh), sandstone (S), siltstone (SS), limestone (LS), and dolomite (DM).

Table 3. Summary of evaluation metrics for various models (cross-validation and training results).

Model	Data set	Pr	R	F1	ROC	PRC
RT	CV	0.895	0.896	0.895	0.927	0.832
	TR	0.901	0.902	0.901	0.935	0.84
RF	CV	0.919	0.92	0.919	0.988	0.972
	TR	0.913	0.914	0.913	0.985	0.967
REPT	CV	0.83	0.833	0.831	0.943	0.857
	TR	0.798	0.798	0.798	0.924	0.827
LMT	CV	0.835	0.836	0.835	0.927	0.857
	TR	0.833	0.833	0.833	0.928	0.856
J48	CV	0.848	0.85	0.848	0.897	0.800
	TR	0.837	0.84	0.837	0.891	0.792
HT	CV	0.452	0.552	0.427	0.614	0.42
	TR	0.525	0.538	0.516	0.737	0.555

Table 4. Performance of the random forest model on each lithology class (cross-validation results).

Class	Pr	R	F1	ROC	PRC
Sh	0.894	0.823	0.857	0.979	0.939
S	0.919	0.95	0.934	0.987	0.983
SS	0.92	0.92	0.92	0.995	0.973
LS	0.927	0.94	0.933	0.997	0.981
DM	0.976	0.993	0.985	1	0.996
Average	0.919	0.92	0.919	0.988	0.972

5.2. Confusion matrix

To evaluate the performance of various models in classifying lithofacies, a confusion matrix was employed. This matrix offers a detailed breakdown of classification accuracy, presenting the percentage of correctly identified instances for each lithology class. Additionally, it provides insight into misclassifications, revealing cases where specific lithofacies were incorrectly assigned to other categories. The optimal technique-derived confusion matrix for the lithologic classes is presented in Table 5.

Evaluation of the confusion matrix (Table 5) revealed varying performance in classifying data points. Notably, the model performs well at identifying instances belonging to the Sh class, achieving high precision (low false positive rate) across all models. This translates to a high level of confidence in Sh predictions, with most being accurate. Conversely, the model exhibited a significant challenge with the DM class, reflected by a high false negative rate across all models. This indicates a frequent misclassification of DM instances, assigning them to other categories. Furthermore, differentiation between LS and DM presented a particular difficulty, especially in models LMT and RT. These findings suggest a need for further investigation to enhance classifier performance, particularly for classes DM and LS. In essence, the models demonstrated mixed success in data point classification. While the Sh class received exceptional treatment, the model struggled with DM and distinguishing between LS and DM.

Table 5. Performance evaluation of optimized classifiers using confusion matrices (cross-validation) a – HT, b – J48, c – LMT, d – RT, e – RF, f – REPT.

Actual label	Sh	1.57	24.29	0.08	0.16	0.58
	S	1.14	50.59	0.15	0.08	0.24
	SS	0.32	7.55	0.11	0.09	0.20
	LS	1.24	4.62	0.19	3.69	1.76
	DM	0.22	4.94	0.21	2.17	3.31
		Sh	S	SS	LS	DM
		Predicted label (a)				
Actual label	Sh	22.84	3.33	0.34	0.62	0.02
	S	2.28	49.52	0.52	0.02	0.01
	SS	0.56	1.08	6.55	0.15	0.03
	LS	0.36	0.01	0.12	11.08	0.10
	DM	0.04	0.03	0.04	0.11	1.13
		Sh	S	SS	LS	DM
		Predicted label (b)				
Actual label	Sh	21.73	3.82	0.36	0.76	0.00
	S	2.88	48.67	0.62	0.01	0.01
	SS	0.64	1.27	6.12	0.19	0.05
	LS	0.52	0.03	0.15	10.73	0.07
	DM	0.04	0.02	0.06	0.14	1.09
		Sh	S	SS	LS	DM
		Predicted label (c)				
Actual label	Sh	17.77	4.77	0.94	0.81	0.07
	S	4.50	41.57	1.46	0.10	0.07
	SS	0.98	1.57	4.71	0.21	0.08
	LS	0.77	0.14	0.24	9.17	0.17
	DM	0.07	0.06	0.06	0.08	9.64
		Sh	S	SS	LS	DM
		Predicted label (d)				
Actual label	Sh	21.36	4.25	0.25	0.80	0.00
	S	1.95	49.93	0.31	0.00	0.01
	SS	0.83	1.76	5.45	0.21	0.02
	LS	0.25	0.07	0.09	11.00	0.09
	DM	0.07	0.04	0.03	0.10	1.11
		Sh	S	SS	LS	DM
		Predicted label (e)				
Actual label	Sh	19.35	3.96	0.31	0.73	0.00
	S	2.69	44.47	0.50	0.01	0.02
	SS	0.74	1.34	5.28	0.15	0.05
	LS	0.38	0.05	0.12	9.85	0.10
	DM	0.02	0.03	0.01	0.03	9.80
		Sh	S	SS	LS	DM
		Predicted label (f)				

5.3. Boosting-based approach

To improve model performance, we utilized ensemble learning techniques, specifically focusing on AdaBoost meta-learners in conjunction with classification tree models. The effectiveness of this approach is evaluated through various performance metrics, including precision, recall, F1-score, area under the precision-recall curve, and area under the ROC curve. These metrics are presented for different models trained and cross-validated, with average and weighted average values provided in Table 6 and Fig. 5, respectively.

The J48 decision tree model fared better than all other models in terms of precision, recall, F1, ROC AUC, and PRC, as shown in Table 6 on both the training set and cross-validation sets. This suggests that J48 performs a better job of accurately classifying events into the relevant groupings. The J48 models performed better than the LMT, REPT, and RF models in the validation and training sets, but their metrics were a little behind. The HT model, on the other hand, performed substantially worse than the other models, particularly in the cross-validation set. This pattern indicates that the HT model may have overfitted the training data. Similarly, in both the training and validation sets, the RT model performed worse on a variety of criteria, particularly the PRC indicator. The assessment findings clearly show that the J48 decision tree model performs better overall across all metrics and data sets. This indicates that J48 is effective in correctly categorizing data points into the appropriate groups.

Fig. 5 illustrates the performance of various ensemble models utilizing boosting with different base learners, including REPT, LMT, J48, and HT. The results suggest that boosting generally leads to improved performance metrics. Notably, boosting with the J48 decision tree classifier achieved the most favorable outcomes. Conversely, the combination of AdaBoost with the RT classifier yielded inferior results compared to other ensemble configurations. It is important to acknowledge that some models exhibited limited responsiveness or even negative reactions to the application of performance-enhancing techniques. To assess the predictive capabilities of the models, three datasets (Set-1, Set-2, and Set-3) were employed. Fig. 6 depicts the prediction accuracy achieved by different algorithms on these diverse datasets.

Based on analysis, there was no significant difference in accuracy between Set-3 and Set-1. Across both datasets, the J48 model performed best, followed by the LMT model. On the other hand, the HT model consistently exhibited the lowest accuracy. For all models, Set-3 yielded the highest accuracy. Furthermore, J48, LMT, REPT, and RF consistently performed across various datasets. However, the HT model remained the least accurate across all three datasets. The J48 decision tree model consistently achieved the highest accuracy in all three datasets. Hence, J48 may be a good choice for these datasets

Table 6. Boosting model performance (cross-validation and training).

Model	Data set	Pr	R	F1	ROC	PRC
HT	TR	0.469	0.547	0.436	0.584	0.389
	CV	0.475	0.541	0.444	0.616	0.405
J48	TR	0.897	0.898	0.897	0.978	0.953
	CV	0.91	0.911	0.91	0.981	0.959
LMT	TR	0.89	0.891	0.89	0.975	0.944
	CV	0.882	0.883	0.882	0.971	0.944
REPT	TR	0.883	0.885	0.883	0.972	0.942
	CV	0.886	0.888	0.886	0.974	0.943
RF	TR	0.892	0.893	0.89	0.981	0.958
	CV	0.888	0.888	0.886	0.979	0.958
RT	TR	0.823	0.822	0.823	0.869	0.738
	CV	0.828	0.829	0.828	0.872	0.743

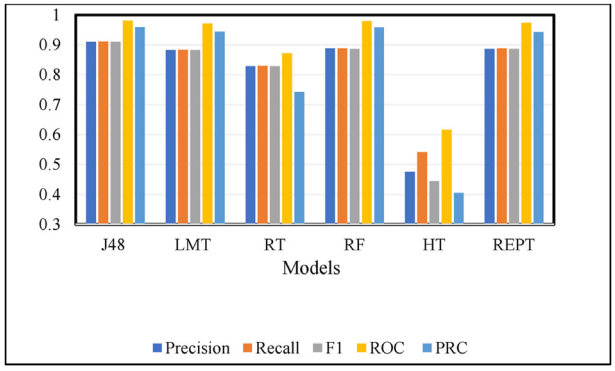

Fig. 5. A description of all algorithm's performance metrics-Boosting method.

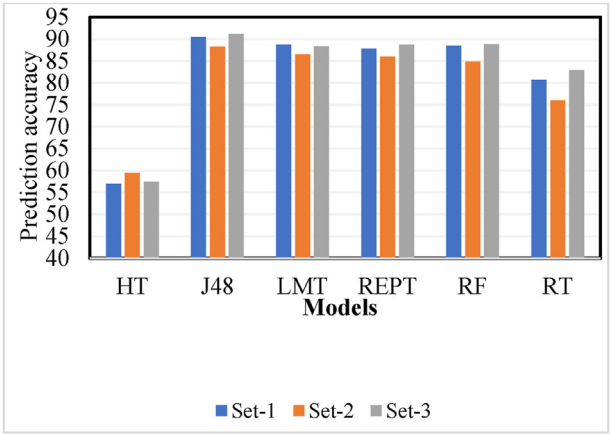

Fig. 6. Prediction of the performance of different models using various datasets.

and prediction tasks. On Set-3, all models except HT are more accurate than on Set-1. Set-3 appears to be easier to learn from, perhaps because it has a clearer structure or better-quality data. The HT model consistently has the lowest accuracy across all datasets. This suggests that the HT model is not well-suited for these prediction tasks, perhaps because it is unable to recognize data patterns.

Although the field of machine learning offers promising prospects for accurate and efficient lithology identification, its application presents several noteworthy challenges. These challenges include the presence of outliers and data extremes, imbalanced datasets, inconsistencies in the dimensionality of logging indicators, and the crucial steps of hyperparameter tuning and attribute selection. However, by effectively addressing these limitations, ML techniques can be applied to achieve robust and streamlined lithology identification processes.

Overall, the implications of this research suggest the application of tree-based machine learning l analysis in petroleum exploration and production operations.

5.4. Conclusions

This research underscores the significance of employing advanced machine learning techniques, data preprocessing methods, hyperparameter tuning, and attribute selection to enhance the accuracy and efficiency of lithofacies prediction in reservoir rocks, particularly in the context of the Camal oil field in Yemen. The study evaluated several machine learning models, with the random forest model demonstrating the best performance in lithology identification, achieving precision, recall, and *F*1-score values of 0.913, 0.914, and 0.913, respectively. Random trees and the J48 model followed in performance, while the Hoeffding Tree model showed the least effectiveness. This study investigated the combined effect of boosting techniques and dimensionality reduction on decision tree models for lithofacies prediction. The results demonstrated a significant improvement in model performance when boosting was incorporated with decision trees. Furthermore, the influence of various dimensionality reduction methodologies on prediction accuracy was explored. These findings emphasize the crucial role of model optimization techniques, particularly boosting algorithms and hyperparameter tuning, in enhancing the effectiveness of tree-based models for lithofacies classification. The study employed outlier removal techniques, normalized data, and addressed imbalanced datasets using the SMOTE to enhance model performance and accuracy.

The findings provide valuable insights for the oil and gas industry in optimizing lithology identification processes using tree-based machine learning models. The prospects for the development of the approaches presented in the research paper are promising, with opportunities for further innovation, integration of advanced algorithms, utilization of big data analytics, automation of prediction systems, and interdisciplinary collaboration. These avenues hold the potential to revolutionize lithofacies prediction in the oil and gas industry, leading to more accurate reservoir characterization and improved decision-making processes.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest/Competing interests

The authors that I have no conflicts of interest to disclose regarding the research, authorship, and/or publication of this work.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author contribution

Abbas M. Al-khudafi: conceived the study, designed the methodology, conducted the analysis, and drafted the original manuscript. **Ghareb M. Hamada**: contributed to manuscript review and editing, data collection, conceptualization, and validation. **Abdelrigeeb Al-Gathe**: provided critical review and approved the final manuscript. **Ibrahim A. Farea**: developed the necessary software and performed validation. **Salem O. Baarimah**: contributed to data collection and visualization. All authors have read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge the substantial contribution of the Ministry of Oil and Mineral Resources of the Republic of Yemen in providing critical data for this study.

References

1. Farouk S, Sen S, Ganguli SS, Abuseda H, Debnath A. Petrophysical assessment and permeability modeling utilizing core

data and machine learning approaches – a study from the Badr El Din-1 field, Egypt. *Mar Petrol Geol*. 2021;133:105265.
2. Potekhin DV, Galkin SV. Use of machine learning technology to model the distribution of lithotypes in the Permo-Carboniferous oil deposit of the Usinskoye field. *J Mining Inst*. 2023;259:41–51.
3. Rashid M, Luo M, Ashraf U, et al. Reservoir quality prediction of gas-bearing carbonate sediments in the Qadirpur field: insights from advanced machine learning approaches of SOM and cluster analysis. *Minerals*. 2023;13:29.
4. Rogers SJ, Fang JH, Karr CL, Stanley DA. Determination of lithology from well logs using a neural network. *Am Assoc Petrol Geol Bull*. 1992;76:731–739.
5. Al Ghathe AA, Hamada GM, AlKhudifi AM. Water saturation determination in carbonate reservoirs hybrid artificial intelligence approaches and conventional techniques. Alexandria, Egypt. In: *Paper Presented at the Middle East Oilfield Conference*. 2016.
6. Al-Anazi A, Gates ID. On the capability of support vector machines to classify lithology from well logs. *Nat Resour Res (Paris)*. 2010;19:125–139.
7. Dan M, Wang Z, Yu-Long, et al. Lithological identification of volcanic rocks from SVM well logging data: case study in the eastern depression of Liaohe Basin. *Petrol Explor Dev*. 2015;42:5.
8. Harris JR, Grunsky EC. Predictive lithological mapping of canada's north using random forest classification applied to geophysical and geochemical data. *Comput Geosci*. 2018;80:9–25.
9. Alfarraj M, AlRegib G. Petrophysical property estimation from seismic data using recurrent neural networks. In: *SEG Tech Progr Expand Abst*. 2018:2141–2146.
10. Pham N, Wu X, Zabihi Naeini E. Missing well log prediction using convolutional long short-term memory network. *Geophysics*. 2020;85:WA159–WA171.
11. Mohamed IM, Mohamed S, Mazher I, Chester P. Formation lithology classification: insights into machine learning methods. In: *Proceedings - SPE Annual Technical Conference and Exhibition*. 2019. https://doi.org/10.2118/196096-ms, 196096-ms.
12. Xie Y, Zhu C, Zhou W, Li Z, Liu X, Tu M. Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. *J Petrol Sci Eng*. 2018;160:182–193.
13. Gong K, Zhihui Y, Chen D, Zhu D, Wang W. Investigation on automatic recognition of stratigraphic lithology based on well logging data using ensemble learning algorithm. *Soc Petrol Eng*. 2018;2018:1–11.
14. Mohamed IM, Mohamed S, Mazher I, Chester P. Formation lithology classification: insights into machine learning methods. In: *Proceedings - SPE Annual Technical Conference and Exhibition, 2019-Septe.* 2019 https://doi.org/10.2118/196096-ms.
15. Nanjo T, Tanaka S. Carbonate lithology identification with generative adversarial networks. *Int Petrol Technol Conf 2020*. 2020:1–10. https://doi.org/10.2523/iptc-20226-m.
16. Sun Z, Jiang B, Li X, Li J, Xiao K. A data-driven approach for lithology identification based on parameter-optimized ensemble learning. *Energies*. 2020;13:1–15.
17. Sun J, Chen M, Li Q, Ren L, Dou M, Zhang J. A new method for predicting formation lithology while drilling at horizontal well bit. *J Petrol Sci Eng*. 2021;196:1–12.
18. Xie Y, Zhu C, Hu R, Zhu Z. A coarse-to-fine approach for intelligent logging lithology identification with extremely randomized trees. *Math Geosci*. 2021;53:859–876.
19. Feng R. Lithofacies classification based on a hybrid system of artificial neural networks and hidden Markov models. *Geophys J Int*. 2020;221:1484–1498.
20. Zhong R, Johnson RL, Chen Z. Using machine learning methods to identify coals from drilling and logging-while-drilling LWD data. In: *SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference*. 2019:1–12 (One Petro).
21. Liu Z, Cao J, You J, Chen S, Lu Y, Zhou P. A lithological sequence classification method with well log via SVM-assisted bi-directional GRU-CRF neural network. *J Petrol Sci Eng*. 2021;205:42030812.
22. Zou Y, Chen Y, Deng H. Gradient boosting decision tree for lithology identification with well logs: a case study of Zhaoxian Gold Deposit, Shandong Peninsula, China. *Nat Resour Res (Paris)*. 2021;30:3197–3217.
23. Lu S, Li M, Luo N, et al. Lithology logging recognition technology based on GWO-SVM algorithm. *Math Probl Eng*. 2022: 1640096. https://doi.org/10.1155/2022/1640096.
24. Abid AA, Salih NM, Martyushev DA. Paleoenvironmental evaluation using an integrated microfacies evidence and triangle model diagram: a case study from Khurmala Formation, Northeastern Iraq. *J Mar Sci Eng*. 2023;11:2162.
25. Martyushev DA, Chalova PO, Davoodi S, Ashraf U. Evaluation of facies heterogeneity in reef carbonate reservoirs: a case study from the oil field, Perm Krai, Central-Eastern Russia. *Geoenergy Sci Eng*. 2023;227:211814.
26. Makarian E, Abad ABMN, Manaman NS, et al. An efficient and comprehensive poroelastic analysis of hydrocarbon systems using multiple data sets through laboratory tests and geophysical logs: a case study in an iranian hydrocarbon reservoir. *Carbonates Evaporites*. 2023;38:37.
27. Wang X, Wang M, Zhao C, et al. Reservoir characteristics and controlling factors of the middle-high maturity multiple lithofacies reservoirs of the Lianggaoshan Formation shale strata in the northeastern Sichuan basin, China. Mar Petrol Geol 161:106692. https://doi.org/10.1016/j.marpetgeo.2024.106692.
28. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;20: 832–844.
29. Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn*. 2005;59:161–205.
30. Alatefy S, Abdel Azim R, Alkouh A, Hamada G. Integration of multiple bayesian optimized machine learning techniques and conventional well logs for accurate prediction of porosity in carbonate reservoirs. *Processes*. 2023;11:1339.
31. Chawla NV, Bowyer KW, Hall LO. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16: 321–357.
32. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*. 2020;415:295–316.
33. Claesen M, De Moor B. *Hyperparameter Search in Machine Learning*. 2015 [arXiv preprint arXiv:1502.02127]. https://arxiv.org/abs/1502.02127.