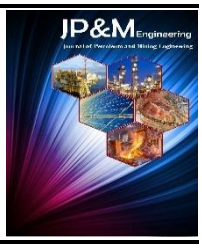
JP&M Engineering
Journal of Petroleum and Mining Engineering



Characterization of Lithfacies Properties of Carbonate Reservoir rocks using Machine Learning Techniques

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

[a]Hadhramout University, Al-Mukalla, Yemen
[b]Arab Academy for Science, Technology & Maritime Transport, Alexandria, Egypt
[c]Emirates International University, Sanaa, Yemen
*Corresponding author e-mail: ghareb.hamada@aast.edu

Article Info

Received 25 Jan. 2024
Revised 1 Mar. 2024
Accepted 13 Mar. 2024

Keywords

Decision tree; Lithology prediction; Machine learning;Well logging; Carbonate Reservoir

Abstract

This study aims to assess the effectiveness of several decision tree machine techniques for identifying formation lithology of complex carbonate reservoir rocks in Gamal oil field. A total of 20966 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, 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 F1-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 prediction capability of models is also carried out using different datasets.

Introduction

In carbonate reservoir rocks, post diagnostic processes such as dissolution, recrystallization, cementation, mineral replacement and dolomitization can bring significant changes in petrophysical properties and then complex mineral pattern and then heterogeneous reservoir. Lithology must be established using well-log data to explore and produce petroleum. The lithology model of a carbonate reservoir rocks 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 of carbonate reservoir rocks. Researchers have recently become more interested in applying machine learning approaches to forecast different types of lithology. These approaches to lithology identification based on machine learning make an effort to train a multi-class classifier model based on a large amount of labelled well-logging data with logging curves, such as gamma ray (GR) 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 first used to classify lithology [3,16]. Support Vector Machine (SVM) was utilized [1] to classify the lithology with logging data points and have accurately identified the lithology facies of heterogeneous sandstone reservoirs. Different types of multi-classification SVM were applied to identify volcanic lithology with well log data [6]. Random Forest (RF) was utilized to predict lithological mapping based on geophysical and geochemical data [9]. In the field of spatial modelling and classification based on log data. Novel hybrid inferential system called ANN-HMM models for lithofacies classification [7]. Approaches to model the rock lithology was developed by using recurrent neural networks were used [2,15].

An artificial neural network model to identify the lithology of a layer as it was being drilled using neighboring well data and real-time drilling data [13] 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, Artificial Neural Network, and Gradient Tree Boosting—were assessed for detection of formation lithology [20].

Conventional single classification algorithms such as decision trees, SVM, and Bayes developed to determine the lithology of the Longqian region of China using well logs [8]. 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 [12]. Generative adversarial networks were presented to recreate thin section images and identify carbonate lithology [14]. An Extreme Gradient Boosting and Bayesian Optimization classifier was proposed for identifying the lithology of the Daniudui and Hangjinqi gas fields [18]. Three machine learning algorithms presented to determine the lithology while drilling. Neural networks (NN), RF, extreme gradient boosting tree (XGBoost) algorithms, and one-versus-one support vector machines (OVO SVMs) are used to create machine learning (ML) [17]. A coarse-to-fine architecture that incorporates outlier detection, multi-class classification, and a tree-based classifier suggested to identify the lithology using two actuals well logging data sets [19]. A hybrid framework consisting of artificial neural networks and hidden Markov models (ANN-HMM) was suggested for the classification of the lithological sequence [11]. They thoroughly evaluated the effectiveness of the suggested classifier using a combination of extreme gradient boosting (XGBoost) and Bayesian optimization (BO) [18]. Coal pay zones were predicted using a variety of machine learning algorithms (LR, SVM, ANN, RF, and XGBoost) and data manipulation methods (NROS and SMOTE) [21]. Bi-directional 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 using the neural networks and hidden Markov models (ANN-HMM) hybrid framework [10]. In contrast to the ANN, SVM, AdaBoost, and RF classifiers, the performance of the gradient boosting decision tree (GBDT) classifier was demonstrated and confirmed [22]. A Gray Wolf Optimization Algorithm (GWO-SVM)-based automatic identification system for lithology logging has been presented [11]. So far it becomes vital to assess the machine learning model's propensity to forecast the kind of lithology under various circumstances. The classification of lithology in the Camal oil field using tree-based machine learning models is tested in this paper using conventional log curves.

Used Well Logging Data

A total of 20966 log data points from four wells were collected from the Camal oil field to perform the evaluation, including seven logging parameters (density log (RHOB), neutron log (NPHI), sonic log (DT), gamma ray log (GR), deep lateral log (LLD), shallow lateral log (LLS), and resistivity log (ML).) with corresponding depths. The output class to be identified is the type of lithology (shale, limestone and dolomite). The range of the seven feature parameters are listed in Table 1. The evaluation was also conducted based on the three datasets listed in Table 2.

Table 1 Range of parameters for lithology classification

Parameter	Maximum	Minimum	STD	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

Table 2 Different datasets used for lithology prediction

Datasets	Points	Input parameters
1	22957	ML, LLD, LLS, GR, RHOB, NBHI, DT
2	32406	GR, RHOB, NBHI, DT
3	20966	ML, LLD, LLS, GR, RHOB, NBHI, DT

Machine Learning Models

Different decision tree-based algorithms for classification approaches were applied. Six typical machine learning models, namely the Random Forest (RF). Random trees (RT), J48, reduced-error pruning decision trees (REPT), logistic model trees (LMT), and Hoeffding Tree (HT). Figure 1 presents the proposed lithology classification methods for this investigation.

Figure 1 Methodology for evaluation ML model.

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.

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 made by each individual tree is returned for regression tasks [23]. The tendency of decision trees to over fit 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. Performance, however, might be impacted by data properties.

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 non-essential 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.

Logistic Model Tree

Combining logistic regression (LR) and decision tree learning, the logistic model tree (LMT) is a classification model that comes with a corresponding supervised training algorithm [24]. 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 [24].

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 exactly the same as that of standard batch learners and runs in sublinear time. It makes use of Hoeffding trees, which take 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.

J48 Classifier

It is a decision tree generation method produced by C4.5. An information-theoretic classification algorithm called C4.5 generates decision trees. Model J48, another name for Ross Quinlan's previous ID3 algorithm. Because C4.5 generates decision trees that are used for classification, the program is frequently referred to as a statistical classifier [4].

Data Preprocessing of Well Logs

Using seven logging features—density log (RHOB), neutron log (NPHI), sonic log (DT), gamma ray log (GR), deep lateral log (LLD), shallow lateral log (LLS), and resistivity log—a total of 20966 well log data points were used for lithology classification (ML).

Outlier Removal

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 remove 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 (Table 3).

Table 3 Prediction accuracy for models utilizing IQR and LOF filters

ML model	IQR	LOF
RT	78.71	79.93
RF	87.57	87.78
J48	81.80	82.72
LMT	81.67	83.51
REPT	78.07	80.15
HT	57.35	59.71

Manage imbalanced dataset

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 [5] specifically to tackle class imbalance issues related to different lithology types, enhancing lithology prediction model performance. Figure 2 shows how the data set's distribution changed both before and after the SMOTE approach was used in various experiments.

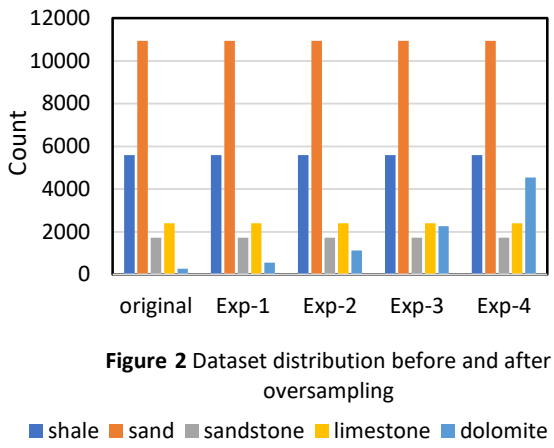


Figure 2 Dataset distribution before and after oversampling

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%.

Normalization

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.

Building Predictive Models

Tree-based models were constructed, namely RF, RT, J48, REPT, LMT, and HT. 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 ten-fold cross-validation technique.

Hyper parameters

Hyper parameters are parameters that control the learning process in machine learning models. unlike other parameters, such as node weights, which are learned during training, hyper parameters are set beforehand [26]. They can be categorized as model hyper parameters, which influence model selection, or algorithm hyper parameters, which affect the learning process's speed and quality. model hyper parameters include factors like neural network topology and size, while algorithm hyper parameters encompass settings like learning rate, batch size, and mini-batch size. Different machine learning algorithms require specific hyper parameters, and tuning them is crucial for adapting models to specific datasets [25]. Tree depth and the total number of trees in a random forest are two instances of hyper parameters for tree models, and learning-related settings like the learning rate, batch size, and mini-batch size.

Tuning Hyper parameter

This study utilized hyper parameter tuning to optimize machine learning models for lithology identification. A 10-fold cross-validation method was employed to find the best hyper parameter set for these models. This approach evaluated the influence of various hyper parameters on model performance and emphasized the importance of hyper parameter tuning in machine learning. The optimal hyper parameters were determined based on the cross-validation results.

In order to construct the model classifier, ten-fold cross-validation was used, and the hyper parameters were optimized. There were ten subsets of the training data in the ten-fold cross-validation process. Nine subsets of the training datasets were chosen for model training and hyper parameter tuning, while one subset was used for model validation. The cross-validation curve for the RT model with hyper parameters is displayed in Figure 3. (K: the number of characteristics selected at random, M: the minimum total weight of leaf instances, and max: the maximum tree depth). From this Figure, the ideal hyper parameters can be found.

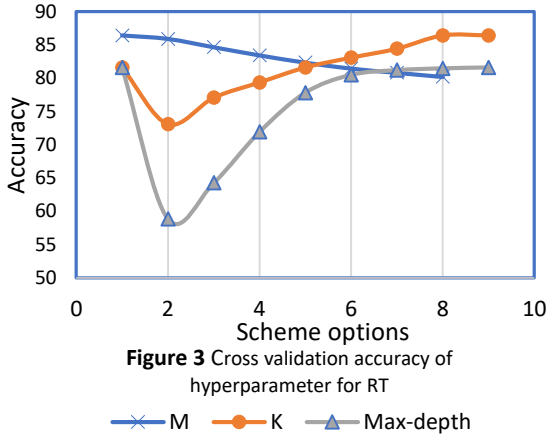


Figure 3 Cross validation accuracy of hyperparameter for RT

Attribute Selection

The correlation between the parameters may affect the ability of the model to forecast. Therefore, the analysis of input data is an important pre-processing technique beneficial for quality control, and through data mining, we can select more influential parameters for lithology detection, reduce the dimension of the input parameter data, reduce overfitting, and improve accuracy. 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 Info Gain, Relief, and One R. The results are presented in Figure 4-6.

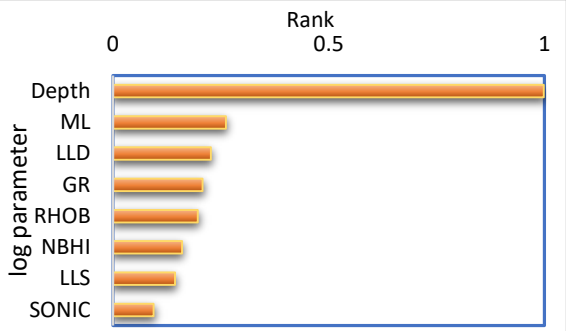


Figure 4 Rank of each feature using InfoGain algorithm.

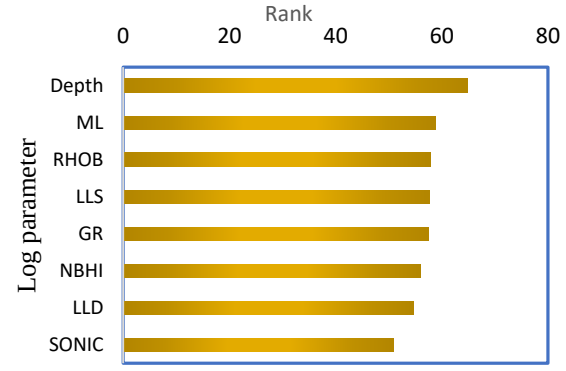


Figure 5 Rank of each feature using OneR algorithm.

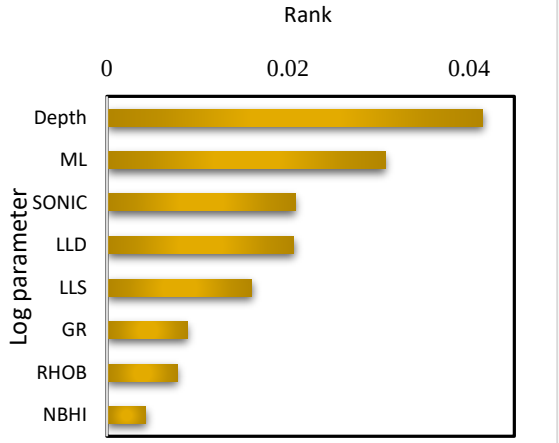


Figure 6 Rank of each feature using Relief algorithm.

To evaluate the prediction capability of the models at various log parameters, classification was performed based on eight (1-8) functional datasets, which are listed in Table 4.

Table 4 Datasets of different functional forms

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

Summary of influence of different type of well logs on performance of different models is presented in Figure 7. Evidently, the degree to which the feature characteristics have an impact is not significantly different for all models except the HT algorithm, which, when considered as a whole, exhibits poor accuracy. All models performed well in Set 5. The RF model provides the best results for each set of variables. Figure 8 shows the link between accuracy for various models and the amount of log parameters.

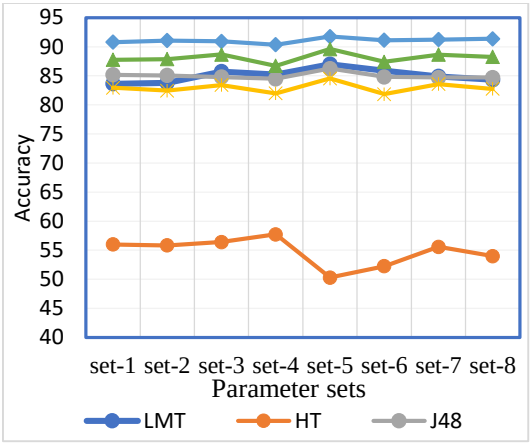


Figure 7 Performance of model with different parameter sets.

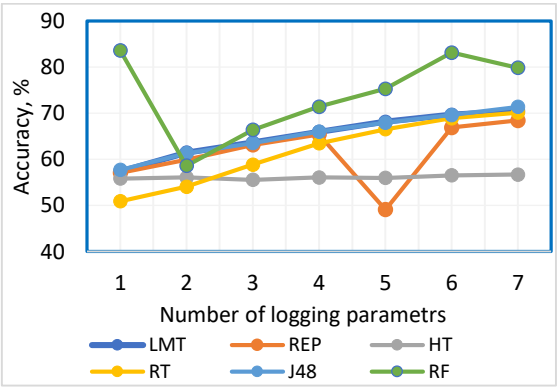


Figure 8 Influence of logging parameters on lithology prediction.

As demonstrated, it has been shown that the log parameters required for lithology prediction can be reduced to four. the recommended methods for identifying lithology include density log (RHOB), neutron log (NPHI), sonic log (DT), and gamma ray log (GR). other logs such as deep lateral log (LLD), shallow lateral log (LLS), and micro resistivity log (ML) are also suggested for this purpose.

Results and discussion

Evaluation metrics and model performance

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 ten-fold cross-validation.

In Table 5, metrics scores for various models are presented.

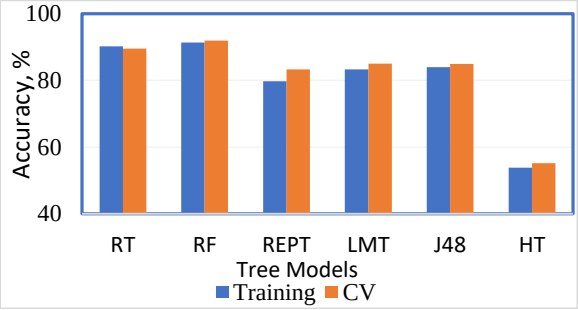


Figure 9 Accuracy for Training and cross validation set of different models.

Table 5 Weighted average of Evaluation metrics for different models- cross validation and Training

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

When compared to the other five classifiers, the RF model produces superior classification results. The HT performs the worst, while the RT is the next-best model by performance. As shown in Figure 9 and Table 6, Among the tree classifiers, the RF has the best performance, with relative high classification accuracy of 92% and 91.35% for cross-validation and training, respectively. This algorithm has high performance in identifying sandstone and dolomite (Table 6). Model J48 ranked third, with average precision, recall, and F-measures of 0.848, 0.85, and 0.848, respectively.

Table 6 Performance of each lithology class -cross-validation for RF model

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
Aver.	0.919	0.92	0.919	0.988	0.972

Confusion matrix

The confusion matrix was utilized to compare the performance of different models in classifying lithology classes. The confusion matrix presents the percentage of correctly classified instances for each lithology class. It highlighted instances where certain lithology classes were incorrectly identified as others. The confusion matrix of the lithological classes, derived for different classification models, is shown in Table 7.

Table 7 Confusion matrix for different classifiers using cross validation
a- HT, b- J48, c- LMT, d- RT, e- RF, f- REPT

Actual label / Predicted label (a)

	Sh	S	SS	LS	DM
Sh	0.19	23.64	0.01	1.66	0.13
S	0.10	47.93	0.02	2.09	0.03
SS	0.01	7.17	0.03	0.70	0.04
LS	0.31	3.64	0.04	6.65	0.40
DM	0.00	1.94	0.05	2.86	0.35

Actual label / Predicted label (b)

	Sh	S	SS	LS	DM
Sh	20.07	4.64	0.80	0.77	0.04
S	3.90	46.41	1.09	0.07	0.05
SS	0.78	1.53	5.59	0.20	0.06
LS	0.50	0.08	0.15	10.47	0.14
DM	0.02	0.06	0.06	0.12	2.41

Actual label / Predicted label (c)

	Sh	S	SS	LS	DM
Sh	18.31	4.37	0.83	0.82	0.03
S	3.95	42.26	1.38	0.05	0.05
SS	0.72	1.44	5.17	0.17	0.05
LS	0.55	0.05	0.19	9.56	0.15
DM	0.03	0.03	0.03	0.05	9.76

Actual label / Predicted label (d)

	Sh	S	SS	LS	DM
Sh	16.66	2.96	0.51	0.58	0.03
S	2.81	36.68	1.06	0.02	0.03
SS	0.47	0.93	11.30	0.12	0.04
LS	0.50	0.03	0.16	8.16	0.10
DM	0.01	0.02	0.03	0.04	16.76

Actual label / Predicted label (e)

	Sh	S	SS	LS	DM
Sh	18.63	3.03	0.37	0.59	0.03
S	1.62	42.11	0.58	0.01	0.03
SS	0.34	0.66	12.92	0.10	0.02
LS	0.26	0.04	0.15	9.18	0.14
DM	0.01	0.01	0.02	0.02	9.14

Actual label / Predicted label (f)

	Sh	S	SS	LS	DM
Sh	13.45	5.19	1.13	0.89	0.08
S	3.38	35.59	1.55	0.02	0.06
SS	0.98	1.90	9.76	0.17	0.06
LS	0.52	0.06	0.23	7.88	0.25
DM	0.04	0.03	0.08	0.07	16.62

Boosting-based approach

Boosting techniques, specifically AdaBoost meta-learners combined with classification tree models, were employed to enhance model performance. Table 8 and Figure 10 present the mean and weighted mean performance measures for classification models employing boosting methodology. These metrics include precision, recall, F1-scores, the area under the precision-recall curve, and the ROC area. The models were trained and cross-validated across different variations.

 DOI: 10.21608/jpme.2024.265484.1190

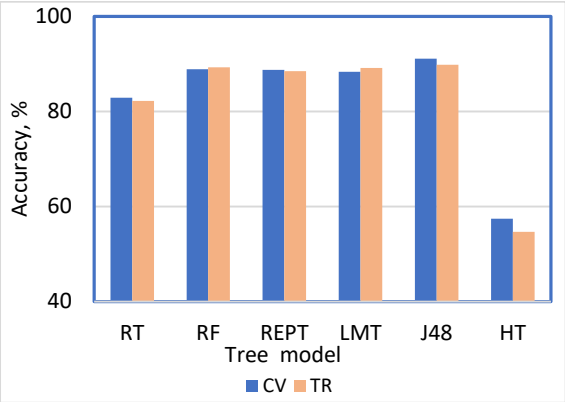


Figure 10 Accuracy for Training and Cross-Validation of a Set of Different Models: A Boosting Approach.

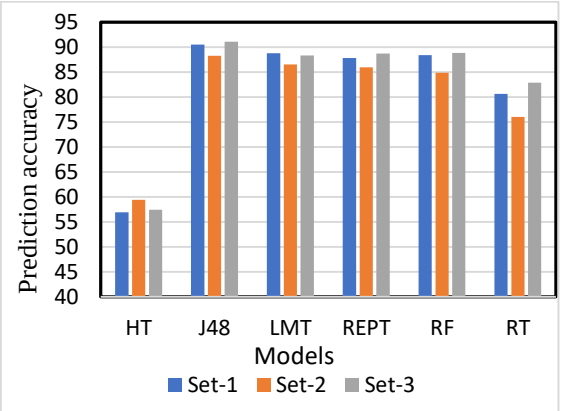


Figure 11 Prediction performance of different models using various datasets.

Tables 10 and Figure 11 provide the average and weighted average performance metrics for classification, such as precision recall, F1-scores, the area of the precision recall curve, and the ROC area, for various models that were trained and cross-validated.

Table 8 Performance metrics for different models- cross validation and training -boosting approach

Model	Data Set	PR	R	F1	ROC
HT	TR	0.469	0.547	0.436	0.584
	CV	0.475	0.541	0.444	0.616
J48	TR	0.897	0.898	0.897	0.978
	CV	0.91	0.911	0.91	0.981
LMT	TR	0.89	0.891	0.89	0.975
	CV	0.882	0.883	0.882	0.971
REPT	TR	0.883	0.885	0.883	0.972
	CV	0.886	0.888	0.886	0.974
RF	TR	0.892	0.893	0.89	0.981
	CV	0.888	0.888	0.886	0.979
RT	TR	0.823	0.822	0.823	0.869
	CV	0.828	0.829	0.828	0.872

The results indicated that using a boosting approach with REPT, LMT, J48, and HT provided good performance metrics. Generally speaking, boosting with J48 provides the best performance metrics. Conversely, combining AdaBoost with the RT classifier yielded poorer results compared to other combinations. Some models showed limited responsiveness or negative responses to performance-enhancing techniques. To assess the predictive performance of the models, diverse data sets were employed for evaluation. Figure 6 demonstrates the precision attained by distinct algorithms across varying databases.

The prediction capabilities of the models were evaluated using different datasets. Figure 11 shows the prediction accuracy of different algorithms based on various datasets, as listed in Table 2.

As shown, all the models provided slightly the same accuracy for set-3 and set-1. As can be seen, the J48 model outperformed all models for both datasets, followed by the LMT, whereas the HT model had the worst performance. The best accuracy was obtained using data set-3 from all the models. Similar results for different datasets were obtained by J48, LMT, REPT, and RF. The lowest accuracy was obtained by the HT model for all the three datasets. The confusion matrix of the lithological classes, derived with an optimal approach, is shown in Table 9.

Table 9 Confusion matrix for different optimized classifiers using cross validation.

a- HT, b- j48, c- LMT, d- RT, e- RF

Actual label	Sh	S	SS	LS	DM
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

Predicted label (a)

Actual label	Sh	S	SS	LS	DM
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

Predicted label (b)

Actual label	Sh	S	SS	LS	DM
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

Predicted label (c)

Actual label	Sh	S	SS	LS	DM
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

Predicted label (d)

Table 10 displays the performance matrices for the top models, J48.

Table 10 Accuracy measures for J48 model for cross-validation

Class	Precision	Recall	F1	ROC	PRC
Sh	0.876	0.839	0.857	0.97	0.927
S	0.915	0.946	0.93	0.98	0.976
SS	0.869	0.762	0.812	0.979	0.883
LS	0.924	0.952	0.938	0.996	0.976
DM	0.987	0.992	0.989	1	0.997
Average	0.91	0.911	0.91	0.981	0.959

As shown in the Table 10, J48 model performed well in distinguishing between limestone and dolomite. Among the models used to categorize limestone and dolomite, J48 performed the best.

The analysis reveals that the models exhibited comparable accuracy levels for set-3 and set-1. In particular, the J48 model demonstrated superior performance across both datasets, followed by the LMT model, while the HT model showed the lowest performance. The highest accuracy was achieved with set-3 across all models. Consistently, J48, LMT, REPT, and RF yielded similar results across various datasets. In contrast, the HT

Conclusion

Some conclusions can be summarized from this work as follows:

1. The study evaluated the effectiveness of various decision tree-based machine learning algorithms for lithology identification using three sets of logging data. The outcomes of training, cross-validation, and experimentation were considered.

2. The Random Forest model demonstrated superior performance in lithology identification, achieving precision, recall, and F-score values of 0.913, 0.914, and 0.913 respectively. The study emphasized the significance of optimization techniques like combining the SMOTE technique with machine learning, hyper-parameter tuning, and boosting-based approaches to enhance model performance.

3. The boosting approach applied to decision trees significantly improved model performance. Furthermore, the study investigated the impact of reducing the dimensionality of input log parameters on model prediction performance using various techniques.

Funding sources

This research was conducted without external funding.

Conflicts of interest

This study was designed and executed free from any bias or external influence.

References

[1] Al-Anazi, A., & Gates, I. D. (2010), On the capability of support vector machines to classify lithology from well logs. Natural Resources Research, 19(2), 125–139. https://doi.org/10.1007/s11053-010-9118-9.

[2] Alfarraj, M., & AlRegib, G. (2018), Petrophysical property estimation from seismic data using recurrent neural networks. SEG Technical Program Expanded Abstracts, 2141–2146. https://doi.org/10.1190/segam2018-2995752.1

[3] Al Ghathe, A.A., Hamada, G.M., and AlKhudifi, A.M., 2016, Water Saturation Determination in Carbonate Reservoirs Hybrid Artificial Intelligence Approaches and Conventional Techniques, MOC, April 19-21, Alexandria, Egypt.

[4] Alatefy S. Abdel Azim R., Alkouh A. and Hamada G., 2023, Integration of Multiple Bayesian Optimized Machine Learning Techniques and Conventional Well Logs for Accurate Prediction of Porosity in Carbonate Reservoirs, Processes 2023, 11, 1339. https://doi.org/10.3390/pr11051339

[5] Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling Technique. 16, 321–357.

[6] MOU, D., WANG, Z., HUANG, Y., XU, S., & ZHOU, D. (2015). Lithological identification of volcanic rocks from SVM well logging data: Case study in the eastern depression of Liaohe Basin. Chinese Journal of Geophysics, 58(5), 1785-1793. doi: 10.6038/cjg20150528.

[7] Feng, R. (2020). Lithofacies classification based on a hybrid system of artificial neural networks and hidden Markov models. Geophysical Journal International, 221(3), 1484–1498. https://doi.org/10.1093/gji/ggaa083

[8] Gong, K., Ye, Z., Chen, D., Zhu, D., & Wang, W. (2018), Investigation on automatic recognition of stratigraphic lithology based on well logging data using ensemble learning algorithm. Society of Petroleum Engineers - SPE Asia Pacific Oil and Gas Conference and Exhibition 2018, APOGCE 2018, 1–11. https://doi.org/10.2118/192006-ms

[9] Harris, J. R., & Grunsky, E. C. (2015), Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Computers and Geosciences, 80(October 2018), 9–25. https://doi.org/10.1016/j.cageo.2015.03.013

[10] Liu, Z., Cao, J., You, J., Chen, S., Lu, Y., & Zhou, P. (2021), A lithological sequence classification method with well log via SVM-assisted bi-directional GRU-CRF neural network. Journal of Petroleum Science and Engineering, 205(42030812), 108913. https://doi.org/10.1016/j.petrol.2021.108913

[11] Lu, S., Li, M., Luo, N., He, W., He, X., Gan, C., & Deng, R. (2022), Lithology Logging Recognition Technology Based on GWO-SVM Algorithm.

Mathematical Problems in Engineering, 2022. https://doi.org/10.1155/2022/1640096

[12] Mohamed, I. M., Mohamed, S., Mazher, I., & Chester, P. (2019), Formation lithology classification: Insights into machine learning methods. Proceedings - SPE Annual Technical Conference and Exhibition, 2019-Septe. https://doi.org/10.2118/196096-ms

[13] Mohammad Ali, A. M. (2015), Artificial Intelligence for Lithology Identification through Real-Time Drilling Data. Journal of Earth Science & Climatic Change, 06(03), 3–6. https://doi.org/10.4172/2157-7617.1000265.

[14] Nanjo, T., & Tanaka, S. (2020), Carbonate lithology identification wth generative adversarial networks. International Petroleum Technology Conference 2020, IPTC 2020. https://doi.org/10.2523/iptc-20226-ms

[15] Pham, N., Wu, X., & Zabihi Naeini, E. (2020), Missing well log prediction using convolutional long short-term memory network. Geophysics, 85(4), WA159–WA171. https://doi.org/10.1190/geo2019-0282.1.

[16] Rogers, S. J., Fang, J. H., Karr, C. L., & Stanley, D. A. (1992), Determination of lithology from well logs using a neural network. In American Association of Petroleum Geologists Bulletin (Vol. 76, Issue 5, pp. 731–739). https://doi.org/10.1306/bdff88bc-1718-11d7-8645000102c1865d.

[17] Sun, J., Chen, M., Li, Q., Ren, L., Dou, M., & Zhang, J. (2021), A new method for predicting formation lithology while drilling at horizontal well bit. Journal of Petroleum Science and Engineering, 196. https://doi.org/10.1016/j.petrol.2020.107955

[18] Sun, Z., Jiang, B., Li, X., Li, J., & Xiao, K. (2020), A data-driven approach for lithology identification based on parameter-optimized ensemble learning. Energies, 13(15), 1–15. https://doi.org/10.3390/en13153903

[19] Xie, Y., Zhu, C., Hu, R., & Zhu, Z. (2021). A Coarse-to-Fine Approach for Intelligent Logging Lithology Identification with Extremely Randomized Trees. Mathematical Geosciences, 53(5), 859–876. https://doi.org/10.1007/s11004-020-09885-y

[20] Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018), Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. Journal of Petroleum Science and Engineering, 160, 182–193. https://doi.org/10.1016/j.petrol.2017.10.028.

[21] Zhong, R., Johnson, R. L., & Chen, Z. (2019), Using machine learning methods to identify coals from drilling and logging-while-drilling LWD data. SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference 2019, APUR 2019. https://doi.org/10.15530/ap-urtec-2019-198288

[22] Zou, Y., Chen, Y., & Deng, H. (2021), Gradient Boosting Decision Tree for Lithology Identification with Well Logs: A Case Study of Zhaoxian Gold Deposit, Shandong Peninsula, China. Natural Resources Research, 30(5), 3197–3217. https://doi.org/10.1007/s11053-021-09894-6

[23] Ho, Kam, Tin. 1998, “The Random Subspace Method for Constructing Decision Forests.” Ieee Transactions on Pattern Analysis and Machine Intelligence 20(8):832–44.

[24] Landwehr, Niels, Mark Hall, and Eibe Frank. 2005, “Logistic Model Trees.” Machine Learning 59(1–2):161–205.

[25] Claesen, Marc, and Bart De Moor. 2015, MIC 2015: The XI Metaheuristics International Conference in Agadir, Morocco

[26] Yang, Li, and Abdallah Shami. 2020, “On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice.” Neurocomputing 415:295–316.