LithoBot: An AutoML approach to identify lithofacies.

Mayur Nawal, Santosh Kumar, Bharath Shekar; Department of Earth Sciences, Indian Institute of Technology, Bombay*

Summary:

Lithofacies identification is a complex problem involving many different challenges such as null values, low accuracy and missing logs. Previously proposed solutions have low levels of accuracy for different lithofacies, need continuous tinkering of hyperparameters and expert understanding of machine learning concepts. We propose LithoBot, a one-click approach to identify lithofacies from well logs. Our framework inculcates numerous combinations of preprocessing, modeling and post modeling analysis within itself and still gives real time, accurate and situation adaptive solutions. It also has an inbuilt module to provide feedback to the end user on how they can improve the lithofacies identification accuracy and plan ahead prior to a wireline log run. The approach is causal and invariant to null values or outliers in well logs. We use open source data from the FORCE-2020 machine learning contest and have achieved state of the art classification accuracy of 94.5%. Automating the prediction along with achieving a higher accuracy opens up the possibility to create similar end-to-end solutions for problems in seismic, well-logging and interpretation domains.

Introduction:

Lithofacies refers to the distinctive physical and chemical properties of rocks (Lee, 2018), allowing for broad categories such as clastic sediments, carbonates, evaporites, or mixtures thereof to be identified (Douglas 1991). A robust lithofacies model derived from sedimentology and lithology data is the basis for higher level stratigraphic and palaeogeographic interpretations (Lee 2018). Generally, methods to identify lithofacies are either laboratory based like particle size analysis, clast lithological analysis, trace element geochemistry, thin section micromorphology (Lee 2018) or are based on the information available from various well logs studied together. Lithofacies have been identified based on signatures observed in Gamma ray (GR) logs like cylindrical, funnel, bell, serrated and sawtooth patterns (Obundun and Mathew, 2012; Nazeer et al., 2016; Nwagwu et al., 2020; Yemets et al., 2021). Other well logs commonly used for lithofacies identification include spontaneous potential (SP), caliper (CALI), density (RHOB), sonic (DT), porosity (NPHI) and photoelectric absorption factor (PEF) (Ma 2019). Recent times have seen significant interest in using machine learning algorithms for lithofacies identification due to their ability to manage large volumes of data, prevent human errors, handle complex conditions and beat human accuracy levels in certain instances. There have been several attempts to use machine learning and deep learning based methods for lithofacies identification. Liu et al. (2020) used support vector machines and multi-kernel learning, Dixit et al. (2020) used probabilistic learning models, Moradi (2021) used k-nearest neighbor algorithm and naive bayes classification. Merembayev et al. (2021) applied wavelet transform to wireline logs in a supervised machine learning setup to identify 12 lithofacies. Their machine learning pipeline consisted of random forest and XGBoost methods. Although these methods have shown promising results, there are certain challenges that need to be addressed. The limitations of the previously mentioned approaches include non-real time lithofacies identification due to their use of acausal rolling windows prior to the application of fourier or wavelet transforms. Further, there is a lack of clarity in addressing missing values in the dataset. Additionally, it is desirable for lithofacies identification algorithms to be simple and easy to use for seamless integration between software and users.

To address the above challenges, we have created an automatic machine learning (AutoML) model. Automatic Machine Learning (AutoML) refers to end-to-end machine learning models that encompass all the steps like pre-processing, feature selection, hyperparameter tuning, model selection, and interpretation in a single framework. We first describe the methodology adopted for the AutoML model designed for lithofacies identification, followed by its application on the data. We find that AutoML can yield robust and reliable results with higher accuracy.

Methodology:

We address the various challenges that come in the way of lithofacies identification with a four-stage algorithm.

Firstly, raw data commonly occur in certain combinations, based on the commonly produced log outputs from the open hole wireline tools (WT) and other logs related to the drilling and mud logging segment. Four commonly occuring combinations, their respective tools and the

GitHub link: https://github.com/mayurnawal123/Lithobot concerned logs are listed in Table 1. The combinations correspond to most of the practical scenarios that may occur during a wireline logging operation. These combinations are not exhaustive and new combinations of logs can be analyzed by the LithoBot framework. Further, the combinations of logs used during training *does not* limit the applicability of LithoBot for prediction with new combinations of logs. Any new log combinations are automatically mapped to the closest model out of the combinations used during training, enabling LithoBot to perform real-time classification. The mapping is achieved based on the best match between new combination log headers and model log headers.

Secondly, although it is common practice to remove any faulty tool data before submission to the client, it is possible that the data is erroneous in certain intervals. LithoBot includes a thresholding algorithm to mitigate the effect of erroneous data on the predictions. Log values beyond the lower and upper threshold for all the logs are converted into null values. The threshold values are taken from the petrophysical limits of the rock properties for each tool and their log values. This step enhances the robustness of the automation, thus reducing concerns around faulty tool data.

The third stage of the algorithm consists of a suite of random forest classifiers that are trained on data from each combination of tools shown in Table 1.

Combination Name	Tool Name	Log Name with units
Combination-1 (C1)	Gamma	GR (API), SGR (API)
	DEN	RHOB (g/cm3) , CALI (in), DCAL (in), DRHO(g/cm3), PEF(barns/e)
	MICROPAD	RXO (ohm-m), RMIC (ohm-m)
	POR	NPHI(%)
	MudLog	MUDWEIGHT (lbs/gallon)
	Others	SP(mV), BS(in)
Combination-2 (C2)	All tools from C1	All logs from C1
	SON	DTC (us/ft), DTS (us/ft)
Combination-3 (C3)	All tools from C2	All logs from C2
	RES	RDEP (ohm-m), RMED (ohm-m), RSHA (ohm-m)
Combination-4 (C4)	All tools from C3	All logs from C3
	DrillData	ROP (ft/hr), ROPA (ft/hr)

Table1: Four combinations of tools and logs that commonly occur together in industrial applications.

Lastly, the weights of the trained models are stored to enable real-time predictions on test data during the application stage. Grouping the data available for prediction into different combinations and mapping the best suitable model ensures that the accuracy is not significantly compromised even with a limited number of logs. LithoBot also helps in future planning of logging sections for improved lithofacies identification. Further, it can suggest additional logs that can help improve classification accuracy and also reports logs that have insufficient data, necessitating reruns of the concerned tools. LithoBot's AutoML approach to identify lithofacies is divided into the following two sub algorithms: one for training and another for application during test time. The classes for the algorithms are schematically represented in Figure 1.

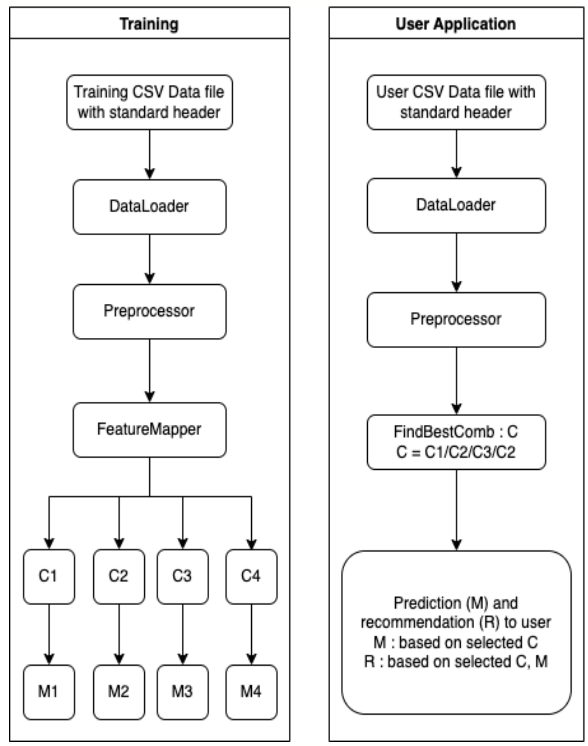

Figure1: Training and user application workflow

Data

We illustrate the performance of LithoBot on well log data from 118 wells released under the NOLD 2.0 license by the Norwegian government and used for the FORCE-2020 machine learning competition (Bormann et al., 2020). The log data acquired from the Norwegian sea comprise more than 1.3 million data points over 14 logs with around 44% null values. The data is imbalanced and certain lithofacies such as shale and sandstone are over-represented as compared to others such as anhydrite, basement etc.. The wells span the south and north Viking graben, with north Viking graben dominated by deeply buried Brent delta facies while south Viking graben has highly variable Permian evaporite geology. The data has been cleaned and despiked. The lithology and lithofacies labels corresponding to the training data have been handcrafted based on a semi subjective interpretation of the well log data. Some of the labels may be prone to errors or variability due to user interpretation.

Results:

Accuracy scores for the 4 combinations are presented in Table 2. The results are obtained on a 80-20 train-test split, with k-fold cross validation (k=3). An accuracy score of 92 % is obtained with the combination C1 formed by the least number of logs (13 logs) , and the score is further enhanced to 94.5 % with the combination C4 that consists of 20 logs. It is clear that additional well logs have non-redundant information that help constrain lithofacies identification. However certain logs may be more important to identify particular lithofacies and the importance of individual logs within combinations is analyzed in a later subsection.

Combination	**C1**	**C2**	**C3**	**C4**
Accuracy	**92**	**93**	**94**	**94.5**

Table 2: Accuracy scores for various combinations.

Detailed results for lithofacies classification for all the combinations and normalized confusion matrix of results for combination-1 are shown in Table 3 and Table 4, respectively. Sandstone/shale has some tendency to be incorrectly labeled as either shale or sandstone, owing to similar RDEP, SP and PEF log values.

	C1		**C2**		**C3**		**C4**	
Lithofacies (L)	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Anhydrite (A)	**96**	**92**	**96**	**95**	**98**	**96**	**97**	**97**
Basement (B)	**95**	**100**	**95**	**100**	**95**	**100**	**95**	**100**
Chalk(C)	**95**	**94**	**96**	**94**	**97**	**95**	**97**	**96**
Coal(C*)	**88**	**62**	**87**	**64**	**88**	**68**	**88**	**70**
Dolomite(D)	**80**	**35**	**84**	**36**	**88**	**39**	**88**	**41**
Halite(H)	**99**	**99**	**99**	**99**	**100**	**99**	**100**	**100**
Limestone(L)	**91**	**76**	**92**	**78**	**93**	**80**	**94**	**80**
Marl(M)	**91**	**84**	**92**	**87**	**93**	**88**	**93**	**89**
Sandstone(S)	**91**	**91**	**92**	**91**	**93**	**93**	**93**	**93**
Sand / Shale (SS)	**87**	**83**	**88**	**84**	**90**	**86**	**90**	**87**
Shale(S*)	**94**	**97**	**95**	**98**	**95**	**98**	**96**	**98**
Tuff(T)	**89**	**83**	**90**	**85**	**94**	**87**	**95**	**88**

Table 3: Precision and recall from the test data for all the combinations and all the lithofacies. P is precision, R is recall (in percentage). Scores for all lithofacies follow the trend, C1<=C2<=C3<=C4.

Halite, chalk, basement, anhydrite have distinctive well log signatures, leading to their identification with a high degree of accuracy. Coal, limestone, marl, and shale exhibit distinctive log values that lead to high accuracy scores. However, the score for the previously mentioned lithofacies are not the highest among all classes, owing to overlaps in log values. It shows that dolomite has poorer results as compared to other lithofacies. This is majorly due to the fact that log values have a highly overlapping tendency when lithofacies are dolomite as compared with other lithofacies. Tuff and shale yield similar well log measurements across most of the tools, leading to a high degree of confusion between the two lithofacies.

Figure 2 shows the reduction in accuracy score when a particular log is completely removed from the input data. Removing traditional logs used for lithofacies identification such as GR, NPHI and DTC collectively reduces the accuracy by 40%. It shows that these logs have high importance in lithofacies identification. It also suggests that combinations which have more logs in them are less affected by missing logs as compared to combinations that have lesser logs. However, no single log dominates the prediction, hence it shows that our method is robust in scenarios when well logs have some logs missing. Missing wireline log information is overlapped with feature importance in Figure 2. Resulting logs are suggested as new logs to be considered in future runs.

Table-5 shows change in prediction accuracy with addition of NaN values during training and testing in each log. The addition of NaN values is over and above the average 44% NaN values originally present in the data.

Combination -1												
L	**A**	**B**	**C***	**C**	**D**	**H**	**L**	**M**	**S**	**SS**	**S***	**T**
A	**92**	**0**	**0**	**0**	**2**	**4**	**0**	**0**	**1**	**0**	**0**	**0**
B	**0**	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
C*	**0**	**0**	**94**	**0**	**0**	**0**	**5**	**0**	**0**	**0**	**0**	**0**
C	**0**	**0**	**0**	**62**	**0**	**0**	**0**	**0**	**12**	**3**	**23**	**0**
D	**1**	**0**	**0**	**0**	**35**	**0**	**2**	**2**	**3**	**5**	**52**	**1**
H	**0**	**0**	**0**	**0**	**0**	**99**	**0**	**0**	**0**	**0**	**0**	**0**
L	**0**	**0**	**1**	**0**	**0**	**0**	**76**	**2**	**4**	**2**	**14**	**0**
M	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**84**	**1**	**1**	**10**	**0**
S	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**91**	**4**	**4**	**0**
SS	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**5**	**83**	**11**	**0**
S*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**98**	**0**
T	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**16**	**83**

Table 4: Normalized confusion matrix for lithofacies identification using log combination C1 (see Table 1).

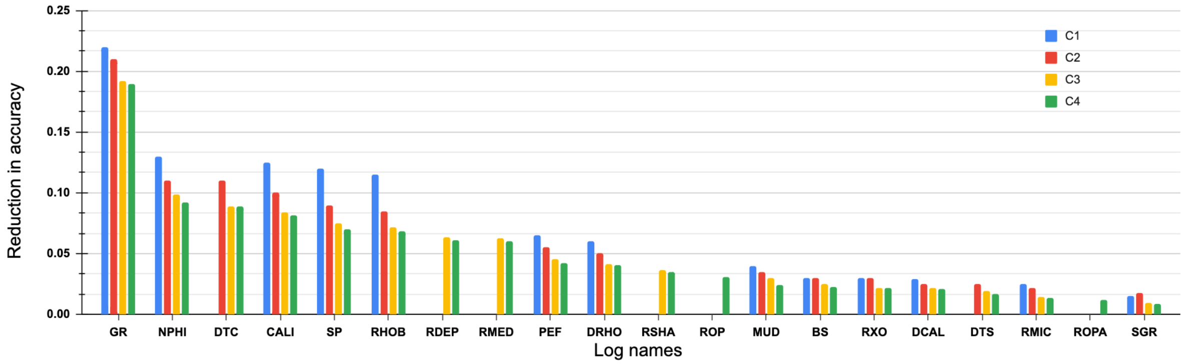

Figure 2: Feature importance plot. The bar values represent the reduction in accuracy corresponding to the scenario when a particular log is missing for a given log combination.

As expected, the higher percentage of missing values lead to a drop in the classification accuracy. However, the classification scores are above acceptable values, suggesting a high degree of robustness in the algorithm.

Discussion:

Our decision to select the random forest algorithm is based on its similarity with human methods (if/else condition on different log values) used for lithofacies identification. We avoided predicting missing values or correcting outliers as lithofacies can be identified using many different log combinations. Penalty matrix that penalizes wrong predictions in a weighted manner can help improve the results for classes that performed relatively poorly, for instance dolomite. Petrophysically, dolomite is distinguished using PEF values (PEF ~ 4), however our data set has this boundary blurred with shale/sand (PEF ~ 2).

Additional % of Nan values	C1	C2	C3	C4
2	91	91	92	93
5	88	89	90	90
8	85	86	87	88
10	83	84	85	86
12	81	82	84	84
14	79	80	82	83
16	78	79	81	81
18	77	77	79	80
20	75	76	78	78

Table 5: Robustness matrix with F1-scores as a function of log combinations and percentage of additional null values in the data.

This leads to dolomite getting wrongly predicted as shale, the dominant lithofacies in the dataset. Table 5 also suggests that with proper selection of training data points, we can train the network with significantly less data.

It should be noted that LithoBot is not restricted to the four combinations of well logs considered above, and can be suitably generalized to other log combinations from wireline logging practice. Tree based deep learning methods can also help in generating greater accuracy of predictions. There is a strong case for working around data redundancy and data binning can help in this direction.

Conclusion:

We have created LithoBot, a real time lithofacies identification framework that works in sync with limited wireline data as it becomes available. The AutoML approach underlying LithoBot integrates all the processing steps including missing values, outlier data, model selection, prediction and feedback on improving results. We achieved an overall accuracy of 94.5 % on data from the FORCE 2020 competition. LithoBot is user friendly and does not require a machine learning background for its operation. The algorithm can aid real time decision making and in planning coring interval decisions by operators. LithoBot can suggest additional logs or rerun of tools to enhance lithofacies prediction accuracy. The approach presented here paves the way for future AutoML solutions in geoscience.