Prediction of photoelectric effect log and facies classification in the Panoma gas field

Felipe F. Melo, PUC Rio

Abstract

I present a machine learning application to perform log prediction and facies classification in the Panoma gas field. The training set is composed of two wells without the photoelectric effect (PE) log and one well with missing values. Before predicting the PE log in two wells, I deal with a few missing data. To predict the logs, I perform feature augmentation in the input logs and generate new logs with a polynomial combination and a low pass filter in wavelet domain. Then, I predict the PE log with the random forest algorithm. In this case, the nested cross validation is used to model selection and hyperparameters tuning. The well Churchman Bible is picked as test well and the score of 72% is achieved on the PE prediction. Both predicted logs are aggregated to the input logs and a new feature augmentation is performed. The new training data is generated aggregating features at neighboring depths and with the vertical gradient. For classification, I used the extreme gradient boosting algorithm and the leave two wells out cross validation to model selection and hyperparameter tuning. The score of 58% is achieved in facies classification on the same test set as prediction.

Introduction

Log estimation, or prediction, is a popular technique in geoscience because some well logs are not acquired through the interest area along the well or simply are not acquired (Bader et al., 2019; Pham et al., 2020). Still regarding well logs, facies classification can be performed in well logs cutoffs (Ma and Zhang, 2019, pp. 232) or from outcomes of seismic inversion (Pendrel and Schouten, 2020), for example. Historically, the knowledge-driven method, where empirical or theoretical knowledge is used to predict a log or classify facies, has being used (Greenberg and Castagna, 1992; Allo, 2019). Alternatively, the data-driven method, i.e., machine learning, which builds computational models that use inference and pattern recognition instead of explicit sets of rules, has being used for a while (Baldwin et al., 1990; Wong et al., 1995) and its popularity increases every day (Dramsch, 2020).

In the literature, there are several works where well logs are predicted with machine learning techniques, for instance: Huang et al. (1996) predicted the permeability, Eshkalak et al. (2014) predicted geomechanical logs for an unconventional reservoir, Akinnikawe et al. (2018) predicted the photoelectric effect (PE) log and Gupta et al. (2019) predicted the sonic (Vp and Vs) and hardness.

Distinct machine learning techniques are being used for facies classification, for instance: Rogers et al. (1992) used neural networks, Cui et al. (2017) applied the principal component analysis, Nishitsuji and Exley (2019) used support vector machine and deep learning. In addition, Yenwongfai et al. (2019) integrate facies and inversions with machine learning.

Regarding the Panoma gas field, Dubois et al. (2007) used artificial neural networks to perform facies classification. Hall (2016a) used support vector machine and proposed a contest with the Society of Exploration Geophysicists (SEG), then this dataset was highlighted. The results of SEG contest presented in Hall and Hall (2017) show that the top-ranked competitors used extreme gradient boosted trees (xgboost). Since then, works with this dataset has being presented by many authors, e.g.: Bestagini et al. (2017) and Zhang and Zhan (2017) used the xgboost to facies classification. With another approach, Mandal and Rezaee (2019) and Mardani (2020) compared the classification results with distinct algorithms. Mandal and Rezaee (2019) computed three synthetic logs (neutron-porosity, density-porosity, and bulk density) before predict the PE log.

In this work, departing from seven wells in the training set, I impute missing values in the PE log in one well and predict the PE log from two wells. Then, I aggregate these wells to the training data in order to perform facies classification. Distinct feature augmentation techniques are used in both log prediction and facies classification. I used the random forest algorithm (Breiman, 2001) for log prediction and the extreme gradient boosting (Chen and Guestrin, 2016) for facies classification. Focusing on the generalization of the models, the hyperparameters were tuned with the nested and the leave two out cross validation, respectively. The well Churchman Bible is picked as test data, instead of the so used Shankle well. The test well is chosen because it has all facies and a large number of samples related to facies of interest. This scenario makes even more challenger the facies classification, in addition to the small number of samples in the training dataset. The scores of 72% are achieve on the PE prediction and of 58% on the facies classification.

Dataset

The Panoma field, located in southwest Kansas and northwest Oklahoma (Figure 1), constitute the largest gas-producing area in North America (Dubois et al., 2007). The dataset consists of 10 wells labeled in Figures 1 and 2, where the well 'Recruit F9' is a pseudo-well generate to better represent the facies Phylloid-algal bafflestone (Hall, 2016b). There are 4149 samples per well log (features) in the dataset, except by the PE log which has 3232 samples.

Prediction of PE log and facies classification

Most of these wells provide seven features (five well log measurements and two indicator variables) and a facies label at every 0.15 m depth intervals. The available well logs are: gamma ray (GR), resistivity (ILD_log10), photoelectric effect (PE), neutron-density porosity difference (DeltaPHI) and average neutron-density porosity (PHIND). The two geologic constraining variables are: nonmarine/marine indicator (NM_M) and relative position (RELPOS). RELPOS is the position of a sample with respect to the base of its respective NM_M interval. The PE log has missing values in one well and is not available in two others.
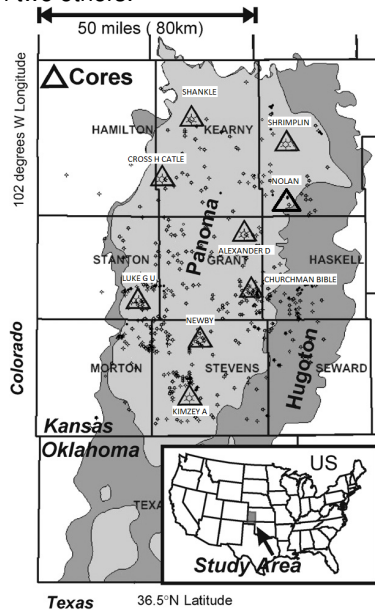

Figure 1 – Map showing the study area, outline of the Panoma field, and nine wells labeled and located by black triangles. The inset shows the location in the central part of the United States of America. Modified after Dubois et. al (2007).

There are nine facies classes (numbered 1–9) identified in the dataset (Dubois et al., 2007; Hall, 2016a) and described in Table 1. Notice that not all of these facies are completely discrete; some gradually blend in to one another. Misclassification of these neighboring facies can be expected to occur. The Adjacent Facies column in Table 1 lists these related classes. Facies 6–8, which correspond to Dolomite (D), Packstone-grainstone (PS) and Phylloid-algal bafflestone (BS), are highlighted in Table 1 because they are the more critical facies since they are the most significant in terms of reservoir storage and flow (Dubois et al., 2007).

Figure 2 shows the distribution of facies for each well, labeled above. Notice that the facies distribution is imbalanced; in fact, not all facies are available in all wells. The last panel shows the distribution of all facies, it's remarkable the imbalance of the facies D and BS, two of the most significant in terms of reservoir.

Figure 3 shows the relationship among all well logs, the samples are colored with the corresponding facies. In the diagonal, a kernel density estimate plot is drawn to show the marginal distribution of the data. Notice that, it is not

clear from these cross-plots the relationship between the log measurements and the facies labels. Therefore, with these data, propose a criterion to facies classification is not straightforward.

Table 1 - Facies labels with their descriptions (Hall, 2016a). The most important facies according to Dubois et al. (2007) are highlighted in green.

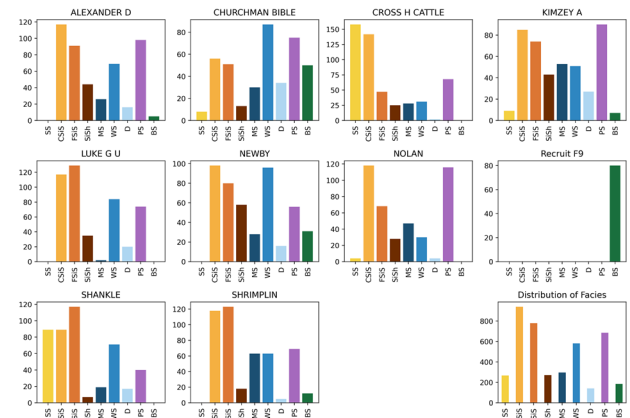| Facies	Description	Label	Adjacent Facies
1	Nonmarine sandstone	SS	2
2	Nonmarine coarse siltstone	CSiS	1,3
3	Nonmarine fine siltstone	FSiS	2
4	Marine siltstone and shale	SiSh	5
5	Mudstone	MS	4,6
6	Wackestone	WS	5,7,8
7	Dolomite	D	6,8
8	Packstone-grainstone	PS	6,7,9
9	Phylloid-algal bafflestone	BS	7,8

Figure 2 – Distribution of facies for each well, the well names are shown over each plot. The last panel shows the distribution of all facies.

Method

This work is divided in two parts: first, the photoelectric effect (PE) log is imputed for a well and computed for other two wells, then these wells are added to the training dataset to perform the facies classification. Both steps are preceded by distinct feature augmentation techniques.

The well Churchman Bible is used as test dataset, i.e., the quality of the regressor (the prediction of the PE log) and of the classifier (the facies classification) will be evaluated on this well. All other nine wells are used as training dataset for facies classification, totaling 2840 samples per feature.
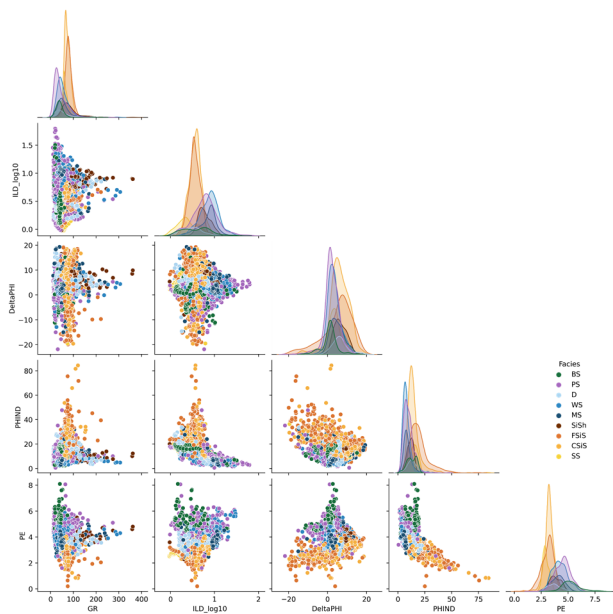
Figure 3 – Cross-plots from the well logs colored with facies classification.

The well Recruit F9 has 12 missing values for the PE log, so I performed data imputation assigning the value of the mean of the samples to the missing ones. The wells Alexander D and Kimzey A don't have the PE log; therefore, these logs are predicted using the Random Forest regressor (Breiman, 2001, Pedregosa et al., 2011), which is an ensemble of multiple decision trees.

Before log prediction, is necessary to perform feature augmentation, which generates new features from the available ones. In this case, it adds non-linearities computing quadratic expansion and by considering all second order interaction terms as in Bestagini et al. (2017). In addition, a low pass filter with the 1D discrete wavelet transform (Mallat, 1999; Souza-Feliciano et al., 2018; Lee et al., 2019) is applied in the logs. The transformation from depth to wavelet domain was performed with the wavelet 'Coiflets 4' and the coefficients were filtered with the soft threshold. Therefore, the training set moves from six to twenty features before estimate the PE logs. Notice that the logs NM_N and RELPOS are not used in feature augmentation because it would make no geophysical sense.

Due to the small training dataset (seven wells, with each of the six features containing 2840 samples), I used the nested cross-validation (CV) for hyperparameter tuning and model selection (Varma and Simon, 2006; Cawley and Talbot, 2010). Nested CV estimates the generalization error of the underlying model and its hyperparameter search. It shows a low bias in practice where reserving data for independent test sets is not feasible (Raschka, 2018). According to Raschka and Mirjalili (2017), in nested CV, there is an outer k-fold cross-validation loop to split the data into training and test folds, and an inner loop to select the model using k-fold cross-validation on the training fold. After model selection, the test fold is then used to evaluate the model performance. Specifically, I used a 7 x 2 cross-

validation and a model with a Random Forest regressor (Pedregosa et al., 2011) was defined.

The feature importance is computed to check the valid of the augmented features. Feature importance can be measured by looking at how much the score decreases when a feature from the validation set is permuted.

After prediction of the PE logs into the two missing wells, these wells are grouped with the seven wells used as training dataset to compound the new training set for facies classification.

Before facies classification I performed a new feature augmentation, distinct from the one for regression. Here, I use two distinct operations: I compute the vertical gradient as in Bestagini et al. (2017) and include a small window to aggregate features at neighboring depths as Lukas Mosser and Alfredo de la Fuente in Hall (2016b) and Hall and Hall (2017). Therefore, the features are augmented from seven to twenty-eight to perform facies classification.

To perform facies classification, I use the Extreme Gradient Boosting (Chen and Guestrin, 2016), which is an optimized gradient boosting (Friedman, 2001). This technique combines many weak learning models together to create a strong predictive model (Raschka and Mirjalili, 2017). It works by sequentially adding predictors to an ensemble, each one correcting its predecessor, i.e., it tries to fit the new predictor to the residual errors made by the previous predictor (Géron, 2019). For cross-validation, I used the Leave 2 Groups Out cross-validator (Hall, 2016b), so all wells were used in the training set except by two that were used in the validation set. These two wells were iteratively update, allowing twenty-eight possibilities for generating a better model.

For the post-processing, I follow the approach from Bestagini et al. (2017). Assuming that facies at neighboring layers are correlated, a median filter is applied to avoid outliers in the facies classification. Thus, the spurious values are replaced by the most present facies in their neighborhood.

The codes used on this work are available at: https://github.com/ffigura/SBGf_ ML_gas/

Results

Figure 4 shows the input well logs, facies and some augmented features to predict the PE logs, from the well Nolan. The logs are: GR, ILD_log10, DeltaPHI, PHIND, PE and Facies, in the upper part, and NM_M and RELPOS, in the lower part. Moreover, the lower part displays some logs generated from feature augmentation: the quadratic expansion generated by the squared GR (GR^2), the second order interaction generated by the multiplication of GR by DeltaPHI (GR x DeltaPHI) and the application of the low pass filter in wavelet domain generated the GR at low frequency (GR low). The feature augmentation produced fourteen logs; for logical reasons, other augmented logs are omitted in Figure 4; however, they are used in the regression part. Notice that to predict the PE logs, the facies logs are not used.

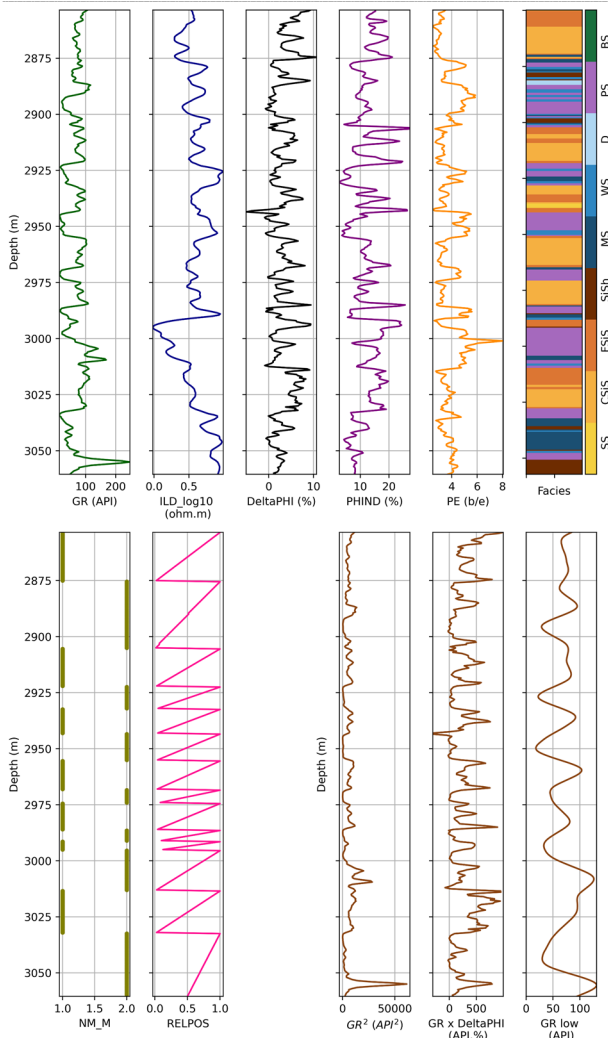# Prediction of PE log and facies classification

Figure 4 – Well logs and facies from the well Nolan. The upper row shows the input logs and facies. The lower row shows two input logs on the left-hand side and three examples of feature augmentation for this well on the right-hand side: the quadratic expansion is shown in the first track with the squared GR, the second order interaction is shown by the GR multiplied by DeltaPHI, and the last track shows the result of the GR after application of the low pass filter in wavelet domain.

Figure 5 shows some results of the prediction of the PE logs. Figure 5a shows the feature importance on the prediction of the test data. As expected, the NM_M produced the most valuable information for the regressor followed by PHIND, then, some augmented features appear before other input logs. The most important augmented features on the test data are: PHIND^2 (the squared PHIND), GR PHIND (the multiplication of GR by PHIND) and ILD_log10 low (the log ILD_log10 after the low pass filter in wavelet domain). This behavior shows the value of the augmented features in the prediction of the PE log. The cross-validation score is 81% with standard deviation of +/- 0.016. Figure 5b shows the true PE log (in black) and the predicted (in orange) on the well Nolan, used in the training set. Figure 5c shows the true (black)

and the predicted (orange) PE log on the well Churchman Bible, the test well. The good fit shown in Figure 5c can be verified qualitatively by the score of 72% on the test data, i.e., the model is not overfitted.
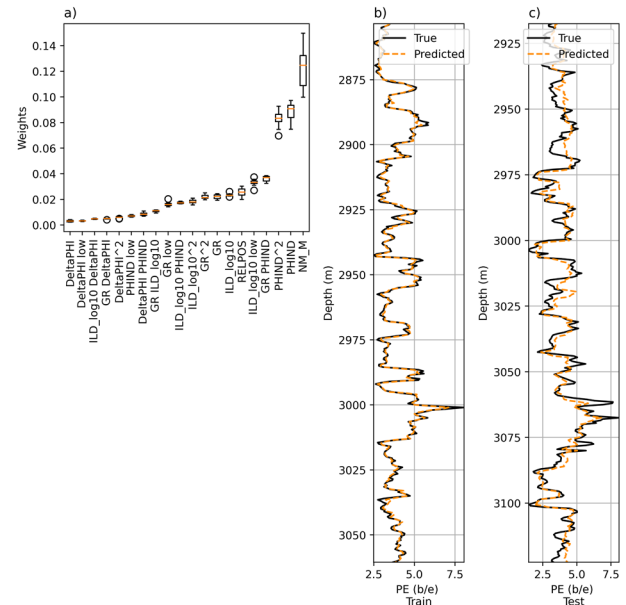

Figure 5 – Prediction results. a) Feature importance for the test set, the mean is in orange. Three augmented features appear with high importance among the five most important. b) True (continuous black line) and predicted (dashed orange line) logs from the well Nolan, this well is part of the training set. c) True (continuous black line) and predicted (dashed orange line) logs from the test set, well Churchman Bible.

Figure 6 shows some input logs, the true and the predicted facies for the test set, the well Churchman Bible. The accuracy of the classification on test set is 58% and the accuracy on the classification for the adjacent facies is 86%. I.e., most of the adjacent facies were correctly predicted, so there were no huge gaps in the prediction, the adjacent facies are described in Table 1.

Figure 7 shows the confusion matrix. Most of the unsatisfactory results are related to adjacent facies. For instance, SS is wrongly predicted as CSiS and FSiS. The same happens with MS, that is mostly wrongly predicted as WS, adjacent facies. The nonmarine facies SS, CSiS and FSiS are mostly wrongly predicted among them, i.e., the NM_M indicator had huge impact in the model. Recall that the facies assigned as BS, PS and D are more critical facies since they are the most significant in terms of reservoir storage and flow. The facies BS have 90% of correct prediction, the facies PS have 57% of correct prediction and the facies D have almost 6% of correct prediction. The well Recruit 9 was generated to better represent the facies Phylloid-algal bafflestone (PS) and probably had an important impact in the prediction because the test set have more PS than any other well, except the well Recruit 9. Notice that, dolomites (D) are not so abundant in the training dataset (last panel on Figure 2). In fact, the test set have more dolomites than any other well

in the training set. Therefore, the model has insufficient data from which to learn a reliable pattern to identify this rock type.
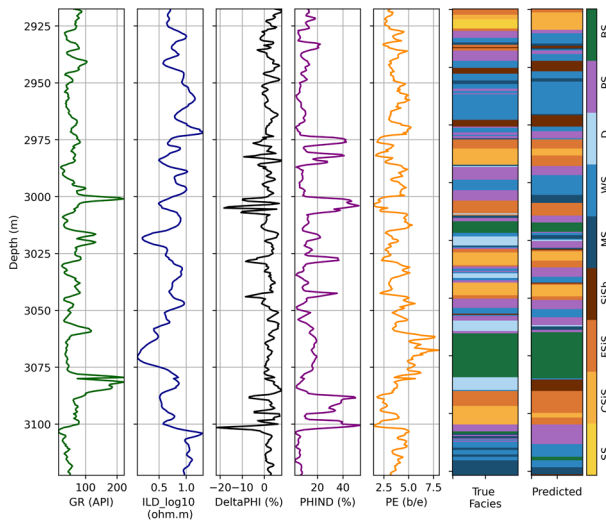

Figure 6 – Facies classification on the test set, the well Churchman Bible. Some input logs with the true and predicted facies. The accuracy of the classification on test set is 58% and the accuracy for the adjacent facies is 86%.
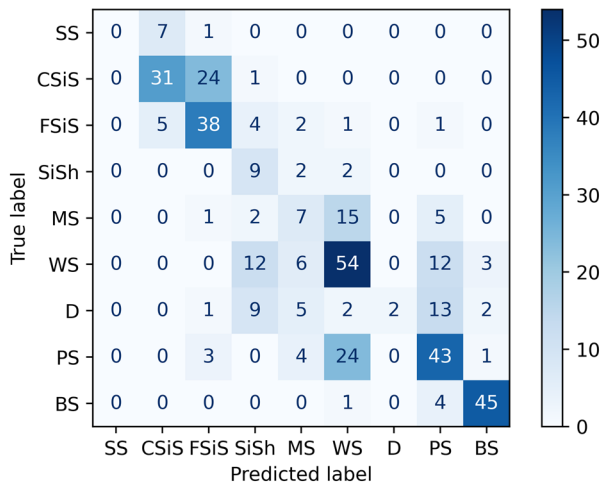

Figure 7 – Confusion matrix. The facies assigned as BS, PS and D are more critical in this work.

Conclusions

I proposed an approach to predict the photoelectric effect log and classify facies in the Panoma gas field. The approach is based on training a model with nine wells and testing the model in the well Churchman Bible. Five measured well logs and two geologic constraining variables are considered as input features. The prediction of the PE log in two wells with a random forest model is preceded by feature augmentation using the quadratic expansion, a second order interaction and the low pass filter in wavelet domain. The facies classification is performed with an extreme gradient boosted trees model. This step was also preceded by feature augmentation, in this case aggregating features at neighboring depths and

with the vertical gradient. The feature augmentation was successful in both prediction and classification. The PE prediction score in the test well shown the good result of this approach. The results for facies classification were acceptable because of the samples limitations. The scores on adjacent facies and on critical facies are good and shown the effectiveness of the proposed approach.

References

Akinnikawe, O.; Lyne, S.; Roberts, J. 2018. Synthetic well log generation using machine learning techniques. *In*: Unconventional Resources Technology Conference. Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers, 1507-1522.

Allo, F. 2019. Consolidating rock-physics classics: A practical take on granular effective medium models. *The Leading Edge*, May 2019, 334-340.

Bader, S.; Wu, X.; Fomel, S. 2019. Missing log data interpolation and semiautomatic seismic well ties using data matching techniques. *Interpretation*, 7, T347–T361.

Baldwin, J.L.; Bateman, R.M.; Wheatley, C.L. 1990. Application of a neural network to the problem of mineral identification from well logs. *The Log Analyst*, 31, 279–293.

Bestagini, P.; Lipari, V.; Tubaro, S. 2017. A machine learning approach to facies classification using well logs. *In:* SEG Technical Program Expanded Abstract, 2137-2142. Society of Exploration Geophysicists.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1), 5-32.

Cawley, G. C.; Talbot, N.L. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.

Chen, T.; Guestrin, C. 2016. Xgboost: A scalable tree boosting system. *In:* Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining, 785-794.

Cui, Y.; Wang, G.; Jones, S.J.; Zhou, Z.; Ran, Y.; Lai, J.; Li, R.; Deng, L. 2017. Prediction of diagenetic facies using well logs – A case study from the upper Triassic Yanchang Formation, Ordos Basin, China. *Marine and Petroleum Geology*, 81, 50-65.

Dramsch, J.S. 2020. 70 years of machine learning in geoscience in review. *Advances in Geophysics*, 61, 1-55.

Dubois, M. K.; Bohling, G. C.; Chakrabarti, S. 2007. Comparison of four approaches to a rock facies classification problem. *Computers & Geosciences*, 33 (5), 599–617.

Prediction of PE log and facies classification

Eshkalak, M.O.; Mohaghegh S.D.; Esmaili, S. 2014. Geomechanical properties of unconventional shale reservoirs. *Journal of Petroleum Engineering*, 10, 961641.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.

Géron, A. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Greenberg, M.L.; Castagna, J.P. 1992. Shear-wave velocity estimation in porous rocks: theoretical formulation, preliminary verification and applications1. *Geophysical prospecting*, 40(2), 195-209.

Gupta, I.; Devegowda, D.; Jayaram, V.; Rai, C.; Sondergeld, C. 2019. Machine learning regressors and their metrics to predict synthetic sonic and mechanical properties. *Interpretation*, 7(3), SF41-SF55.

Hall, B. 2016a Facies classification using machine learning. *The Leading Edge*, 35, 906–909.

Hall, B. 2016b. Facies classification using machine learning: GitHub < https://github.com/seg/2016-ml-contest > (accessed December 19, 2020).

Hall, M.; Hall, B. 2017. Distributed collaborative prediction: Results of the machine learning contest. *The Leading Edge*, 36, 267–269.

Huang, Z.; Shimeld, J.; Williamson, M.; Katsube, J. 1996 Permeability prediction with artificial neural network modeling in the Ventura gas field, offshore eastern Canada. *Geophysics*, 61, 422–436.

Lee, G.R.; Gommers, R.; Wasilewski, F.; Wohlfahrt, K.; O'Leary, A. 2019. PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, 4(36), 1237.

Ma, Y.Z.; Zhang, X. 2019. *Quantitative geosciences: Data analytics, geostatistics, reservoir characterization and modeling*. Springer International Publishing.

Mandal, P.P.; Rezaee, R. 2019. Facies classification with different machine learning algorithm – An efficient artificial intelligence technique for improved classification. *In:* ASEG Extended Abstracts.

Mallat, S. 1999. *A wavelet tour of signal processing*. Academic Press.

Mardani, R.A. 2020. Various Machine Learning Classifiers to Predict Rock Facies. *In:* CSEG GeoConvention abstracts.

Nishitsuji, Y.; Exley, R. 2019. Elastic impedance based facies classification using support vector machine and deep learning. *Geophysical Prospecting*, 67, no. 4, 1040–1054

Pendrel, J.; Schouten, H. 2020. Facies—The drivers for modern inversions. *The Leading Edge*, 39(2), 102-109.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Pham, N.; Wu, X.; Naeini, E. Z. 2020. Missing well log prediction using convolutional long short-term memory network. *Geophysics*, 85(4), WA159-WA171.

Raschka, S.; Mirjalili, V. 2017. *Python machine learning*. Packt Publishing Ltd.

Raschka, S. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808.*

Rogers, S.J.; Fang, J.H.; Karr, C.L.; Stanley, D.A. 1992. Determination of lithology from well logs using a neural network. *AAPG bulletin*, 76(5), 731-739.

Souza-Feliciano, A.C.; Alvarez-Candal, A.; Jiménez-Teja, Y. 2018. Wavelet theory applied to the study of spectra of trans-Neptunian objects. *Astronomy & Astrophysics*, 614, A92.

Varma, S.; Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1), 1-8.

Wong, P.M.; Jian, F.X.; Taggart, I.J. 1995. A critical comparison of neural networks and discriminant analysis in lithofacies, porosity and permeability predictions. *Journal of Petroleum Geology*, 18(2), 191-206.

Yenwongfai, H.; Mondol, N.H.; Lecomte, I.; Faleide, J.I.; Leutcher, J. 2019. Integrating facies-based Bayesian inversion and supervised machine learning for petro-facies characterization in the Snadd Formation of the Goliat Field, south-western Barents Sea. *Geophysical Prospecting*, 67, no. 4, 1020–1039.

Zhang, L.; Zhan, C. 2017. Machine Learning in Rock Facies Classification: An Application of XGBoost. *In:* SEG Global Meeting Abstracts: 1371-1374.