Data mining and machine learning for porosity, saturation, and shear velocity prediction: recent experience and results

Roberto Ruiz[1*], Anna Roubickova[2], Cyrille Reiser[1] and Neelofer Banglawala[2] explore the potential of mining an extensive petrophysics and rock physics well database in the Norwegian Sea through advanced machine learning algorithms for estimation of reservoir elastic properties, and what it could mean for the optimization of petrophysical and rock physics workflows.

Introduction

Building an accurate model of the subsurface is of great importance for the oil and gas industry. The more precise a model is, the lower the risk in exploration for hydrocarbons will be. Companies often leverage a wealth of existing tools and workflows to support this. However, they all rely upon the same fundamental measurements and data sets, for instance well log data.

Well logs are physical measurements collected from drilled boreholes and can provide a very accurate 1D view of the subsurface. As robust as they are, once this raw data is collected, it must be processed by specialists (a petrophysicist, and/or a rock physicist) into a set of conditioned logs containing a detailed description of the mineralogy, porosity, and fluid saturation information as well as elastic logs, compressional wave velocity or Vp, shear wave velocity or Vs, and density or RHOB. This well data and its interpretation can be later integrated with seismic data in Amplitude Versus Offset (AVO) and Quantitative Interpretation (QI) analyses to improve the understanding of identified reservoirs and prospects, or for exploration in frontier areas.

Deriving this set of conditioned elastic logs, fluid and mineral interpretation is a lengthy process, which can take more than one week depending on the geological complexity. Any opportunity to optimize this workflow and integrate regional knowledge is highly valuable. This is where machine learning (ML) can help. The petrophysical/rock physics workflow at its core is a pattern recognition and prediction exercise, which can be posed as an ML problem where there is an input set of curves (raw logs) and an output set of curves (mineralogy, porosity, fluid saturation, conditioned elastic logs).

The implementation of ML techniques in the industry for well log prediction is not new. Bhatt and Helle (1999), implemented neural networks for the prediction of porosity, permeability, and Total Organic Content (TOC) from well logs. Jiang et al. (2020), explored the use of support vector regression (SVR), random forest (RF), and the multilayer perceptron (MLP) for addressing the scale mismatch between seismic and geological layers when predicting porosity logs, whereas Grana et al. (2020), looked at using ML for facies classification. This study investigates whether data mining of existing petrophysical and rock physics

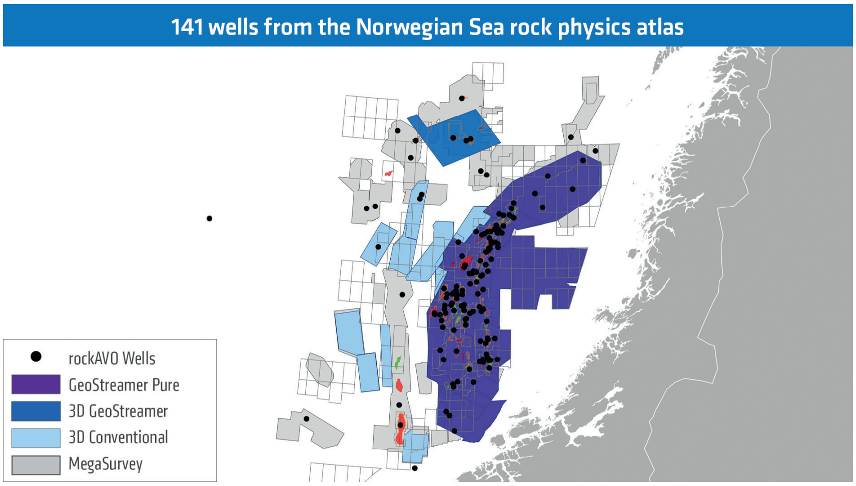

Figure 1 Distribution of 141 exploratory wells used in this study in relation to oil fields (green), gas fields (red) and PGS 3D seismic data in the Norwegian Sea. An additional 190 wildcat wells are available in the region. At one week per well it would take more than three years of petrophysicist's work to condition all wells in the area.

[1] PGS | [2] EPCC
[*] Corresponding author, E-mail: roberto.ruiz@pgs.com

DOI: 10.3997/1365-2397.fb2021055

libraries using ML provides results that are as high quality as those produced via conventional (manual) petrophysical and rock physics workflows. The work focuses on predicting total porosity, hydrocarbon saturation, and Vs, which is often missing in wells and expensive to acquire, but crucial in AVO analysis.

We use data mining of a multi-client rock physics library (PGS rockAVO) composed of 141 wells, sampling a variety of formations and lithologies, ranging from Tertiary to Triassic age in the Norwegian Sea (Figure 1). Each of these wells has previously undergone a thorough petrophysical evaluation, as well as rock physics conditioning. This data set provides an extensive and robust range of observations for training and testing the prediction algorithms.

The robustness and success of the approach demonstrates the potential for integrating existing petrophysical and rock physics libraries with ML for the estimation of porosity, hydrocarbon saturation and Vs in a short time; once a model is trained, the prediction can be performed in seconds. We also show that a porosity model trained in the Norwegian Sea, can be adapted well in another basin.

Theory and workflow

Two different ML algorithms have been used in this study. Parts of the workflow were predicted using a multi-layer perceptron. A perceptron is a mathematical abstraction of a single neuron in a human brain, which can be chained and combined in a layered manner to increase the complexity and predictive power of the model. These models are often referred to as Neural Networks (NN). The models in this study were implemented using the PyTorch library (Paszke et al., 2019), which offers an effective level of control over the configuration of the network and each of its layers, as well as over the learning parameters.

Tree-based models, such as decision trees, RF and Boosted Trees (BT) present an alternative to neural networks. NN models will learn to imitate a functional relationship between the input and target properties. Tree-based models, on the other hand, derive their prediction from the recorded target properties of a group of observations that are similar to the current sample, where the similarity rules are defined during the training process. The work described here uses BT, as these are less prone to overfitting the training data than other tree-based algorithms. The models were implemented using the XGBoost library (Chen and Guestrin, 2016), an open-source software library providing a regularizing gradient boosting framework for many languages.

One key advantage of the BT algorithm over NN is that it handles missing measurements relatively well, whereas NN can only predict where the full suite of input logs is present. Missing measurements at a particular depth is a well-known problem when working with well logs. Sometimes issues arise with the logging tool and a particular log cannot be acquired over a certain interval, and NN can neither train nor predict from an incomplete set of inputs, which significantly reduces the amount of data we can work with. On the other hand, a missing measurement of an important feature may negatively impact the accuracy of the prediction.

Both NN and BT rely on a loss function, a measure that expresses how close the predicted target is to the ground truth. The most used measure is the Mean Squared Error (MSE), which calculates the average square of differences between the true and predicted values across the test or validation set. This provides a good idea of the fit on the level of individual observations, i.e., the depth steps recorded throughout each well. As predictions are realized from individual observations, MSE was chosen as the measure to optimize all the models in the work.

A log of a property through a well is more than a collection of values. There are also trends and changes of direction to consider, which the MSE cannot capture. To overcome this issue, two additional measures were implemented: predictability (PEP) and goodness of fit (R^2).

The predictability score (Silva et al., 2015; White, 1997) measures how much better a set of predictions is compared to predicting a constant 0 for the whole set. A model that precisely predicts the true values scores 1, which is the best possible score.

The R^2 score (Kramer, 2005) is a similar concept, but it measures how much better the predictions are compared to a constant expected value, i.e., the average target value for the given set. A model predicting the expected value would have $R^2=0$, while the best score is 1. Intuitively, R^2 can be interpreted as the proportion of variance of the target variable explained by the input features.

All three measures provide complementary information – MSE helps to assess the general accuracy of the models, while PEP and R^2 help to understand the overall fit of the predictions.

Figure 2 summarizes the main workflow applied for the prediction of each target property in this study. LAS files from 141 wells were loaded as Pandas data frames in a Jupyter notebook. This makes it easier to prepare and feed our models. Given the high quality of the data only a quick clean-up was performed – mainly removal of rows where the target property is not present, and validation of unit consistency across all wells for each curve. Then features were selected. Feature selection is a pivotal point in ML and refers to the choice of properties presented to the model to derive the target property from. We performed a

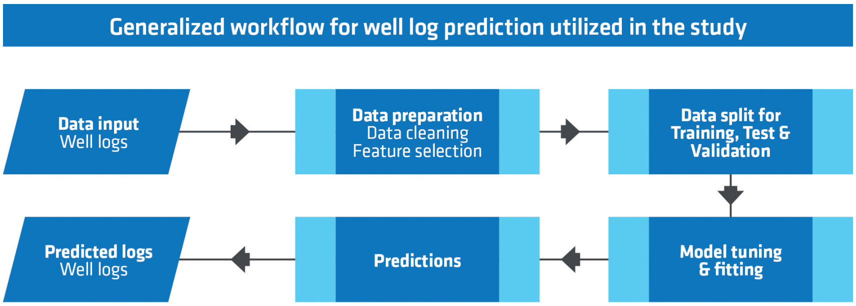

Figure 2 Generalized workflow for well log prediction utilized in the study.

manual filter and removed obviously redundant properties. Then the BT model applied its built-in feature selection as this is embedded in its training process. This is a process similar to what a petrophysicist would perform, who albeit having more than 12 curves available, usually concentrates in a few key logs such as: gamma ray (GR), neutron porosity (NPHI), deep resistivity, RHOB, Vp and Vs.

The data was randomly split on a well-by-well basis, 113 wells (80%) were used for training and 28 wells (20%) were kept aside for blind testing of the model. This ensured that the ML algorithm is predicting on wells that it has never seen, allowing us to estimate the accuracy of the predictions and the generalization power of the model.

To ensure the optimal behaviour of the models, a hyper-parametrization search was performed. Hyper-parametrization refers to setting up the parameters of the model and the training process that remain unchanged during the training, such as maximum allowed tree depth or smallest number of observations required for a node to split. This is a lengthy process, as the algorithm must identify the best combination of model parameters within predefined ranges of the parameters' values. We implemented this search using the Hyperopt (Bergstra et al. 2015) library and the optimal parameters were applied in the final model that predicts the target well log.

Porosity prediction results

Total porosity (PhiT) is the proportion of fluid-filled spaces in the rock, and it can be given as a percentage or fraction of the total space; we use the latter definition. As porosity is a property that cannot be measured directly in the inside of the wellbore, the porosity log must be computed.

In the Norwegian Sea well database, the PhiT log is calculated from bulk density using the mass balance equation. The process requires a mineralogic and fluid saturation interpretation. The interpretation enables an initial estimate of PhiT using empirical trends (Wyllie et al., 1956; Raymer et al., 1980). A rock physicist then takes this initial petrophysical interpretation and PhiT estimation and establishes a relationship between observed velocities or impedances to PhiT, fluid saturations, diagenetic cement, grain sorting and size, mineralogy, etc., through rock physics models in a process known as rock physics diagnostics or RPD (Dvorkin and Nur, 1996). Once the rock physics model and parameters are calibrated to the observed elastic response, it is possible to model the final PhiT. The volume of minerals, fluids saturations and elastic logs are often revisited and updated during this final stage, so that all sets of data (measured logs, petrophysical interpretation, well reports, fluid analysis test, rock physics models, etc.) are consistent, making the estimation of total PhiT a complex and extensive exercise.

Porosity Model	MSE	R2	PEP
Base logs	0.00034	0.95213	0.99416
Processed logs	0.00016	0.97723	0.99722
Minimal logs	0.00011	0.98475	0.99814

Table 1 PhiT prediction accuracy estimation for each model in well 6507/11-7. Different metrics were used to estimate the error fitting, although only MSE was used as loss function for the ML algorithms.

The PhiT prediction was modelled using the BT algorithm, and we studied the impact of different inputs to establish the most suitable sets of features, considering both the accuracy of the predictions as well as availability of the input features. The following sets of inputs were considered:

A. **Base logs -** a set of curves like those used by the petrophysicist, namely: GR; NPHI; logarithm of deep resistivity; raw elastic logs; unconditioned RHOB or RHOB_RAW; unconditioned Vp or Vp_RAW and unconditioned Vs or Vs_RAW. Ramm and Bjørlykke (1994) identified a positive correlation between burial depth and porosity in the area, therefore the true vertical depth below mudline (TVDBML) was also used as an input for the model
B. **Processed logs -** conditioned elastic logs, plus GR, NPHI, logarithm of deep resistivity and TVDBML
C. **Minimal logs -** only conditioned Vp, RHOB and TVDBML

Figure 3 shows the results of all three models for blind test well, 6507/11-7 in the Norwegian Sea. The match between the black (true PhiT) and the red curve (base logs model) is very good, the model seems to be unaffected by the long missing section of the NPHI, raw RHOB and raw Vs in the shallow-to mid-section of the well. The fitting is excellent for the blue curve (predicted from processed logs) as well as for the green curve (porosity predicted from minimal conditioned set of curves).

The different metrics used to estimate the accuracy of each of the models for this well are shown in Table 1. The MSE is

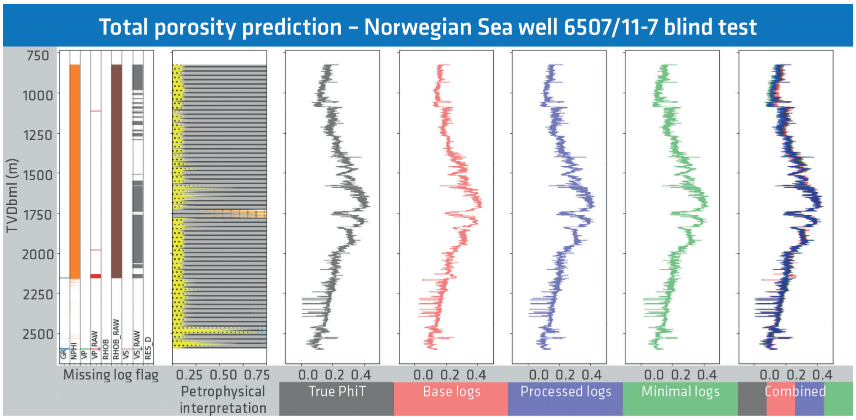

Figure 3 PhiT (fraction) prediction using XGBoost regression algorithm in the Norwegian Sea area with three different suites of logs as inputs. The base model deals with significant sections of missing NPHI, RHOB_RAW and Vs_RAW (shown on the leftmost track) but delivers a robust prediction nevertheless. The petrophysical interpretation is presented for reference although volumetrics are not used as input in the prediction.

very low, but note that this is a squared error so the processed and minimal logs models deviate on average by 1% from the true PhiT, while the average error of the base logs is under 2%. Both PEP and R^2 are very close to 1 for all models, with R^2 more effectively capturing the differences in the models' predictions.

Table 1 also shows that the minimal logs model is the best performer for this well. This might be due to gaps in the RHOB_RAW log that are filled in the conditioned RHOB log by modelling a synthetic density curve from Vp using the rock physics model calibrated during the RPD phase. However, the excellent model fitting where the RHOB log has not been conditioned, indicates that Vp and RHOB logs might carry most of the necessary information for predicting PhiT in the area. More importantly, the results indicate that the model does not require either a mineralogical interpretation, or a fluid saturation estimation, to derive an accurate PhiT for the entire well length, which is a requirement for any conventional methods of estimating this reservoir property.

Geographic transferability test of porosity prediction models

Encouraged by the results, we tested the suitability of these Norwegian Sea porosity prediction models for a well located in a different basin. Figure 4 shows the performance of the models on well 7122/4-1 from the Barents Sea. Although the geological history of this basin is very different, base and minimal log models performed well. This indicates that a ML model trained in one region could be used to estimate an initial PhiT in a new well, from a new area directly from raw logs. A specialist would need to perform a more time-consuming and robust calibration.

Table 2 shows that the overall fit for base and minimal is slightly worse than the Norwegian Sea, though errors remain low.

Porosity Model	MSE	R2	PEP
Base logs	0.00097	0.63504	0.9263
Processed logs	0.00706	-1.65277	0.46439
Minimal logs	0.0005	0.81327	0.96229

Table 2 PhiT prediction metrics for well 7122/4-1 (Barents Sea), although the model has never seen information from this region, the prediction models show a good performance.

Similarly to the previous experiment, the minimal logs model outperforms the other two.

Hydrocarbon saturation prediction results

Hydrocarbon saturation is a key reservoir property. When encountered in a wellbore, fluid saturation can define whether a well, a prospect, a field or a region is an economic success. Therefore, it is critical to have a representative estimation of hydrocarbons in a well. The petrophysicist will focus on a particular reservoir, then look at a series of logs and fluids reports, to choose the most appropriate equation to estimate water saturation.

The most used equations in the oil industry are Archie (1942), Simandoux (1963), and Poupon and Leveaux (1971). Each depend on a series of logs (like porosity or shale mineral fraction) and parameters that are carefully derived and calibrated for a specific reservoir. The hydrocarbon saturation log in the 141 wells in the Norwegian Sea has been estimated primarily using the Archie equation. However, no discrimination in the 80-20% well-by-well split has been performed.

The saturation prediction has been performed using three different methodologies:

1. **HCS1** – using BT to directly predict the saturation of hydrocarbons
2. **HCS2** – using a BT classifier to identify whether there are hydrocarbons at a particular depth (setting up a true/false Boolean flag), then, using a different BT model, an estimation of the hydrocarbon saturation where the hydrocarbon flag was 'true'
3. **HCS3** – using a NN to classify whether there are hydrocarbons at a particular depth, then an estimation of the hydrocarbon content using the BT model

All models were trained using the same set of input logs, namely: TVDBML; GR; NPHI; Vp_RAW; RHOB_RAW and a logarithm of deep resistivity. Figure 5 illustrates the hydrocarbon saturation prediction from each of the models in well 6507/3-10. There are a few missing input logs at the top and bottom of the well, but in the main reservoir section, all input logs are available. All three approaches agree with the true hydrocarbon saturation (calculated using the Archie equation), although models HCS1 (red curve) and HCS3 (green curve) show more residual saturations of hydrocarbon than HCS2 (blue curve). This may be accurate,

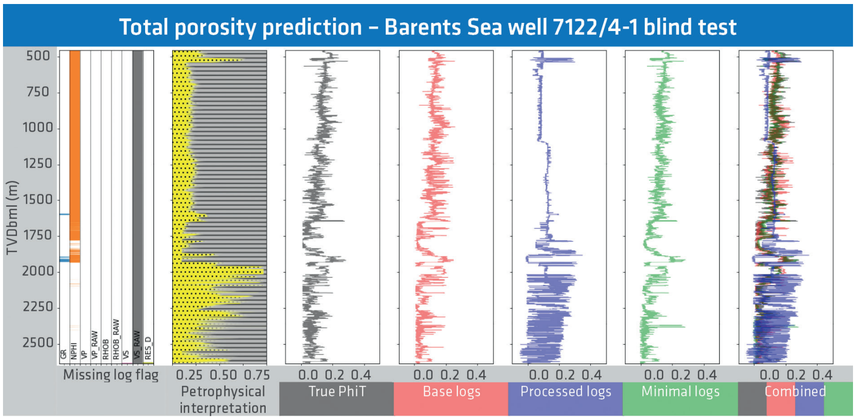

Figure 4 PhiT (fraction) predictions result in a blind well from the Barents Sea. The processed logs model seems to be affected by the conditioning of the logs, while the minimal logs model is the most robust, suggesting that the additional inputs included in the base logs introduce noise rather than useful information.

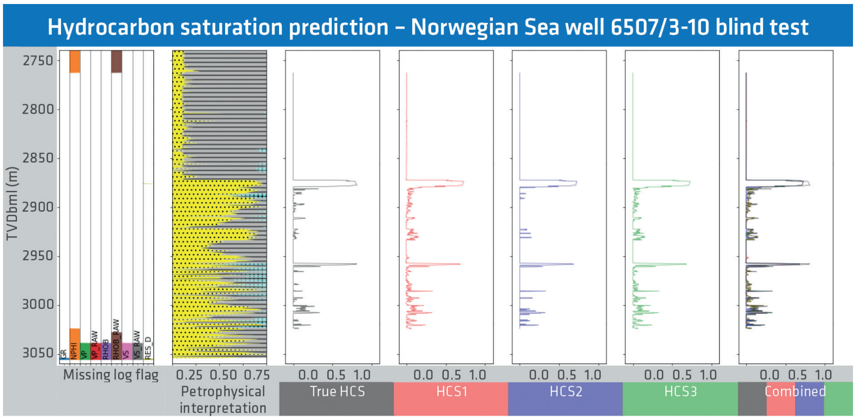

Figure 5 Blind well hydrocarbon saturation (fraction) prediction using three different approaches. All models produce a relatively good prediction, HCS2 is particularly good at handling saturations clipped to zero unlike HCS1 (BT regression only) and HCS3 (NN classifier plus BT regression). All models have been trained using a standard set of curves without any conditioning or processing applied.

SHC Model	MSE	R^2	PEP
HCS1	0.001671	0.822064	0.830401
HCS2	0.001504	0.839939	0.847438
HCS3	0.001604	0.829283	0.837282

Table 3 Hydrocarbon saturation error estimation for different models in well 6507/3-10. MSE was used as loss function for the ML algorithms.

as it is known that petrophysicists clip the residual saturation of hydrocarbon to zero to simplify models.

The error metrics are listed in Table 3. The three models have nearly identical MSE as most of the target values are 0 (which the models frequently correctly identify), and this dominates the error evaluation.

Shear-wave velocity log prediction results

Shear waves are acoustic waves in which particles oscillate perpendicular to the direction the wave propagates. The shear wave carries critical information that, in conjunction with Vp and RHOB logs, provides the framework for performing advanced QI studies using seismic data via AVO analysis (Castagna et al. 1993).

Vs logs are acquired using a full waveform acoustic tool, which generates compressional, shear and Stoneley waves, making the acquisition of this type of log expensive and sparse in old wells. To overcome this, several empirical equations have been proposed to estimate a synthetic Vs log. The common Vs predictors in the oil and gas industry are Greenberg and Castagna (1992) and Krief et al. (1990). Both approaches establish relationships between Vp, porosity, and a reference fluid, to the Vs log. A robust petrophysical interpretation is required prior to the application of these approaches, on top of a careful calibration of the main parameters in the equations, to produce a reliable final synthetic Vs log.

Vs estimation benefits an approach like porosity prediction using BT models:

A. **Base logs** – a model with GR, NPHI, logarithm of deep resistivity, RHOB_RAW, Vp_RAW and TVDBML
B. **Processed logs** – model trained using GR, NPHI, logarithm of deep resistivity, conditioned Vp, conditioned RHOB and TVDBML
C. **Minimal logs** – a model that only considers conditioned Vp, RHOB and TVDBML

Results of models from well 6406/12-2 are illustrated in Figure 6. The first track shows a long section of missing NPHI and Vs_RAW log in the shallow section, but not the conditioned Vs. This indicates that the true Vs log is composed of a measured Vs log and a synthetic, where the flag missing VS_Raw, is true. The synthetic section of the target log has been predicted by petrophysicists using the Greenberg and Castagna predictor with a modified shale coefficient.

The base logs model (red curve in Figure 6) produces a reasonable prediction despite lacking significant information from the NPHI log. This implies the model could be used as a tool to estimate an initial Vs that might be used in early AVO analysis, while the final Vs curve is validated by specialists. The

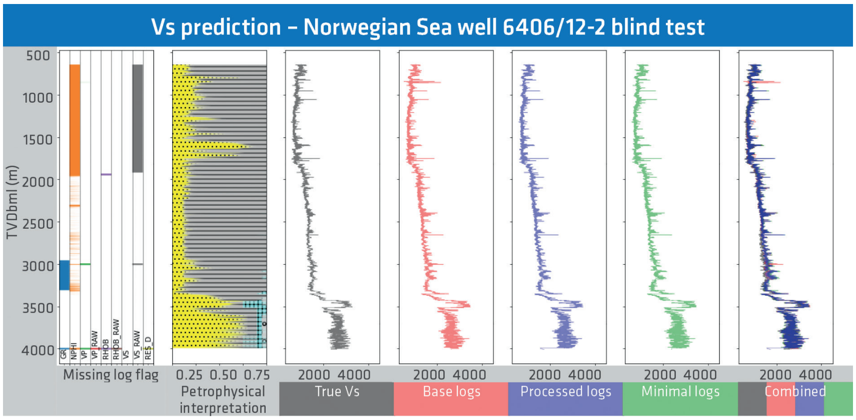

Figure 6 Vs (m/s) prediction using BT algorithm in the Norwegian Sea area with three different suites of logs as inputs. The gap in the true Vs log upper section has been filled by a specialist using Greenberg-Castagna relationship with calibrated coefficient for the shale, the bottom half of the true Vs is measured. The three BT models predict quite well in both zones of the true Vs log (acquired and synthetic). The petrophysical interpretation is presented for reference although volumetrics are not used as input in the prediction.

Vs Model	MSE	R2	PEP
Base logs	8807.99	0.9793	0.99617
Processed logs	10285.6	0.97583	0.99553
Minimal logs	12374.2	0.97092	0.99462

Table 4 Vs prediction errors and fitting metrics for well 6406/12-2 for all models.

other two models, the processed logs (blue curve) and minimal logs model (green curve) also produce accurate predictions. Table 4 shows the overall fitting between the predictions and the true Vs.

Conclusions

Results presented in this paper demonstrate the potential of using ML algorithms to accurately predict porosity, hydrocarbon saturation and Vs from measured well logs, aided by an extensive petrophysical and rock physics atlas in the Norwegian Sea. Unlike traditional empirical approaches, this method does not require inputs such as mineralogy or fluid saturation, to perform a robust prediction.

Prior works attempting similar tasks using analogous techniques (e.g., Hall, 2016; Bestagini et al., 2017) work on much smaller data sets as ML models are hard to scale in the geophysical context due to the noise inherent to large data sets, which may prevent models from discovering true relationships between different features. However, this work demonstrates that some petrophysical properties are consistent across numerous wells, and even across different geographic locations, making ML algorithms a very promising option for estimating properties accurately and efficiently, as well as an extremely useful tool for optimizing current petrophysical and rock physics workflows, along with reducing the overall turnaround. As an example, we observed that it took under 25 minutes for a standard workstation to train our three different porosity models, and only a fraction of a second to predict porosity from these three models on a new well.

Acknowledgements

The authors would like to thank the Impact Acceleration Account Rapid Response (IAA) fund for supporting the development of this work and PGS MultiClient for providing the data (well CPIs and rockAVO Atlas) and technical collaboration.

This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk) funded by the University of Edinburgh and EPSRC (EP/P020267/1).

References

Archie, G.E. [1942]. The electrical resistivity log as an aid in determining some reservoir characteristics. *Transaction American Institute of Mechanical Engineers*, **146**, 54-62.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. and Cox, D.D. [2015]. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science and Discovery*, **8**(1), p.014008.

Bestagini, P., Vincenzo, L. and Tubaro, S. [2017]. A machine learning approach to facies classification using well logs. *SEG Technical Program Expanded Abstracts,* 2137-2142.

Bhatt, A. and Helle, H.B. [1999]. Porosity, Permeability and TOC Prediction from Well Logs Using a Neural Network Approach. *61st EAGE Conference & Exhibition*, Extended Abstract.

Castagna, J.P., Batzle, M.L. and Kan, T.K. [1993]. Rock physics – The link between rock properties and AVO response. In *offset-Dependent Reflectivity – Theory and Practice of AVO analysis, investigations in Geophysics*, No. 8 ed. J.P. Castagna and M. Backus. Tulsa, OK: Society of Exploration Geophysicists, 135-171.

Chen, T. and Guestrin, C. [2016]. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM and SIGKDD international conference on knowledge discovery and data mining*, August, 785-794.

Dvorkin, J. and Nur, A. [1996]. Elasticity of high porosity sandstones: Theory for two North Sea data sets. *Geophysics*, **61**, 1363-1370.

Grana, D., Azevedo, L. and Liu, M. [2020]. A comparison of deep machine learning and Monte Carlo methods for facies classification from seismic data. *Geophysics*, **85**, 41-52.

Greenberg, M.L., and Castagna, J.P. [1992]. Shear-wave velocity estimation in porous rocks: Theoretical formulation, preliminary verification and applications. *Geophysical Prospecting*, **40**, 195-209.

Hall, B. [2016]. Facies classification using machine learning. *The Leading Edge*, **35**, 906-909.

Jiang, L., Castagna, J.P. and Russell, B. [2020]. Porosity prediction using machine learning. *90th SEG Annual Meeting Conference & Exhibition*, Expanded Abstract.

Kramer, M. [2005]. R2 statistics for mixed models. *15th Annual Conference on Applied Statistics in Agriculture.*

Krief, M., Garat, J., Stellingwerff, J., and Ventre, J. [1990]. A petrophysical interpretation using the velocities of P and S waves (full-waveform sonic): *The Log Analyst*, **31**, November, 355-369.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. [2019]. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, **32**, 8024–8035.

Poupon, A., and Leveaux, J. [1971]. Evaluation of water saturation in shaly formations. *Trans. Soc. Prof. Well Log Analysts, 12th Annual symposium*, Paper O.

Ramm, M. and Bjørlykke, K. [1994]. Porosity/depth trends in reservoir sandstones: assessing the quantitative effects of varying pore-pressure, temperature history and mineralogy, Norwegian Shelf data. *Clay Minerals*, **29**, 475-490.

Raymer, L.L., Hunt, E.R. and Gardner, J.S. [1980]. An improved sonic transit time-to-porosity transform. *Society of Petrophysicists and Well Log Analysts SPWLA 21st Annual Logging Symposium,* July, paper.

Silva, E., Seabra, C., de Campos, R. and Augusto, F. [2015]. Impact of Sonic Calibration in the Predictability between a Seismic Trace and its corresponding Reflectivity. *14th International Congress of the Brazilian Geophysical Society*, August 2015.

Simandoux, P. [1963]. Mesures dielectriques en milieu poreux, application a mesure des saturations en eau: Etude du comportement des massifs argileux. *Revue de l'Institut Français du Pétrole,* **18**, supplementary Issue, 193-215.

White, R. [1997]. Accuracy of well ties – practical procedures and examples. *67th SEG Annual Meeting Conference & Exhibition*, Expanded Abstract.

Wyllie, M.R.J., Gregory, A.R. and Gardner, L.W. [1956]. Elastic wave velocities in heterogeneous and porous media. *Geophysics*, **21**, 41-70.