

# Well-Log-Based Reservoir Property Estimation With Machine Learning: A Contest Summary

Lei Fu<sup>1</sup>, Yanxiang Yu<sup>1</sup>, Chicheng Xu<sup>1</sup>, Michael Ashby<sup>1</sup>, Andrew McDonald<sup>1</sup>, Wen Pan<sup>2</sup>, Tianqi Deng<sup>2</sup>, István Szabó<sup>3</sup>, Pál P. Hanzelik<sup>3</sup>, Csilla Kalmár<sup>3</sup>, Saleh Attwah<sup>4</sup>, and Jaehyuk Lee<sup>5</sup>

## ABSTRACT

Well logs are processed and interpreted to estimate in-situ reservoir properties, which are essential for reservoir modeling, reserve estimation, and production forecasting. While the traditional methods are mostly based on multimineral physics or empirical formulae, machine learning provides an alternative data-driven approach that requires much less a-priori geological or petrophysical information. From October 2021 to March 2022, the Petrophysical Data-Driven Analytics Special Interest Group (PDDA SIG) of the Society of Petrophysicists and Well Log Analysts (SPWLA) hosted a machine-learning contest aiming to develop data-driven models for estimating reservoir properties, including shale volume, porosity, and fluid saturation, based on a common set of well logs, including gamma ray, bulk density, neutron

porosity, resistivity, and sonic. Log data from nine wells from the same field, together with the interpreted reservoir properties by petrophysicists, were provided as training data, and five additional wells were provided as blind test data. During the contest, various data-driven models were developed by the contestants to predict the three reservoir properties with the provided training data set. The top five performing models from the contest, on average, beat the performance of the benchmarked Random Forest model by 45% in the root-mean-square error (RMSE) score. In the paper, we will review these top-performing solutions, including their preprocessing techniques, feature engineering, and machine-learning models, and summarize their advantages and conditions.

## INTRODUCTION

Well logs are processed and interpreted to estimate in-situ petrophysical, geomechanical, and geochemical reservoir properties, which are essential for reservoir modeling, reserve estimation, and production forecasting (Xu et al., 2019). The modeling is often based on multimineral physics or empirical formulae (Deng et al., 2020; Luyckx et al., 2022). When a sufficient amount of training data, either from core or logs, is available, and the data are properly organized, a machine-learning solution provides an alternative approach to estimate those reservoir properties based on well-log data and is usually with less turnaround time and human involvements (Khan et al., 2018; Singh et al., 2020; Li et al., 2022). Machine learning has been widely applied to a spectrum of petrophysical task scenarios, including rock classification (Zhang and Zhan, 2017; Xu et al., 2022), petrophysical regression (Anemangely et al., 2019), and image segmentation (Oh et al., 2019).

Following the success of the first SPWLA PDDA SIG Machine-Learning Contest in 2020 (Yu et al., 2021), a second contest was launched with another practical petrophysics task scenario: to develop data-driven models for estimating reservoir properties, such as shale volume, total porosity, and water saturation, based on a common set of well logs, including gamma ray, bulk density, neutron porosity, resistivity, and acoustic slowness. The contest started on November 1, 2021, and ended on February 1, 2022. The contestants were provided the raw properties log data from nine wells from the same field together for the feature set. The target variables are the petrophysical interpretations of shale volume (VSH), porosity (PHIF), and water saturation (SW). These properties were core calibrated by the petrophysicists. In machine learning, the input data used to train models are referred to as features. We will refer to the input log data as features and the output properties as targets throughout when discussing the machine-learning approach. The contest needed to build a data-driven model and develop

Manuscript received by the Editor June 5, 2023; revised manuscript received September 21, 2023; manuscript accepted September 25, 2023.

<sup>1</sup>SPWLA PDDA SIG, Aramco Americas: Aramco Research Center – Houston, lei.fu.rice@gmail.com; Amazon Web Services, yueureka@gmail.com; Aramco Americas: Aramco Research Center – Houston, Chicheng.Xu@aramcoamericas.com; Devon Energy, ashby149@aol.com; Geoactive Limited, andy.mcdonald@geoactive.com

<sup>2</sup>Team UTFE: University of Texas at Austin, wenpan@utexas.edu; tianqizx@utexas.edu

<sup>3</sup>Team MoLPhy: MOL Group, ISzabo3@MOL.hu; hanzelik.pal@gmail.com; kalmarcsilla@gmail.com

<sup>4</sup>Team Attwah\_Analytics: saleh.z.attwah@gmail.com

<sup>5</sup>Team Jaehyuk Lee: Baker Hughes, Jaehyuk.Lee@bakerhughes.com

the machine-learning workflow with the provided training data set, and then they will deploy the newly developed data-driven models on the blind test data set that is composed of five nearby wells to predict the reservoir properties based on the well-log data.

The target reservoir properties used for training and evaluating the models, including porosity, saturation, and shale volume, were based on interpretations provided in the contest data set rather than direct measurements. As such, there is inherent subjectivity and uncertainty in this “ground truth” data due to human and methodological biases inherent in the interpretations. This has implications that will be discussed further regarding the contest results.

The metric to evaluate the performance of the model is the root-mean-square error (RMSE), which is formulated as:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}$$

where  $m$  is the number of samples,  $\hat{y}_i$  is the predicted value, and  $y_i$  is the value interpreted by petrophysicists. Both are vectors with three elements: shale volume, porosity, and water saturation. All three vectors have been normalized to the range between 0 and 1.

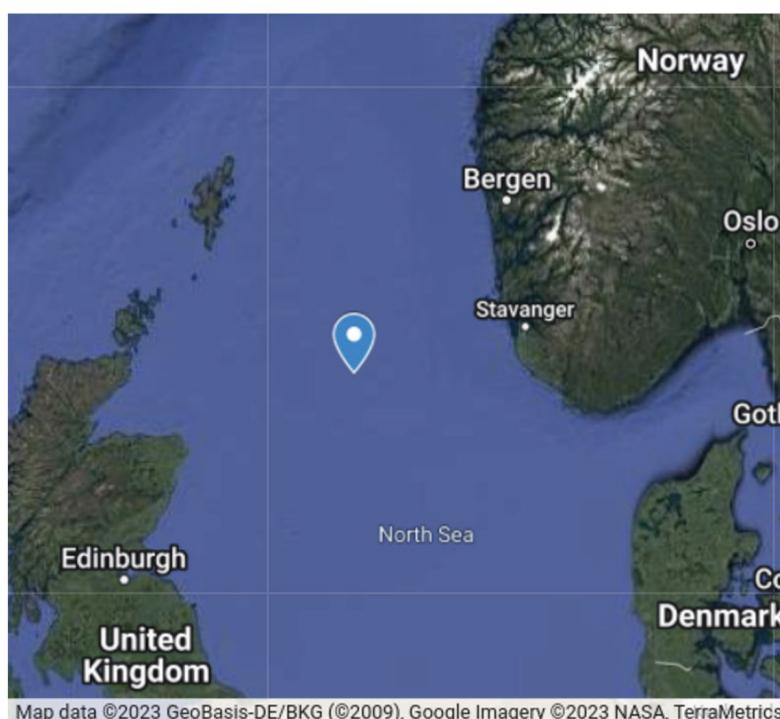
We have also developed a Random Forest model as the benchmark model (Fu et al., 2021), where we only removed missing values and applied default hyperparameters. We

used the default values of the Random Forest provided in scikit-learn. Some of the hyperparameters are listed here: the number of trees in the forest is 100; the function to measure the quality of a split is the mean-squared error; the minimum number of samples required to split an internal node is two; the minimum number of samples required to be at a leaf node is one. The averaged RMSE score of 0.117 was achieved in the benchmark model.

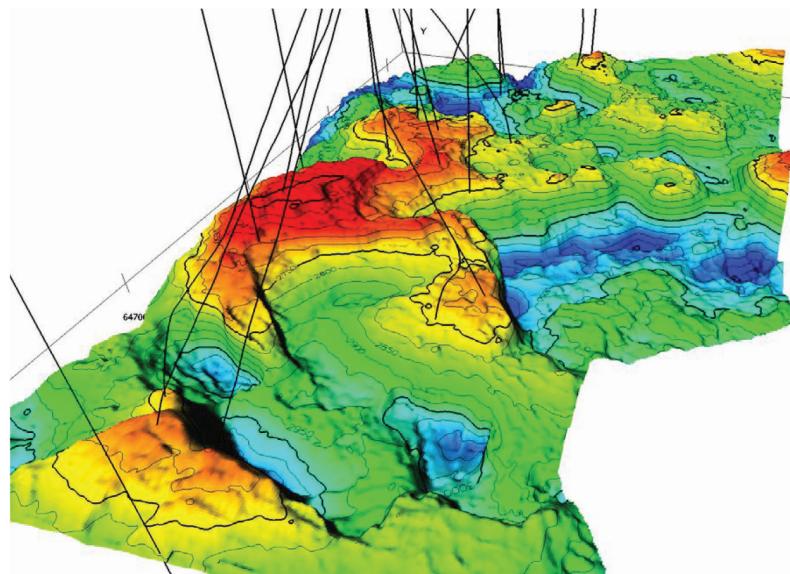
## DATA SET DESCRIPTION

The data set used in this contest is from the Equinor Volve field data (<https://www.equinor.com/en/what-we-do/digitalisation-in-our-dna/volve-field-data-village-download.html>). Equinor officially released a complete set of data from the Volve Field in the North Sea for research, study, and development purposes. Well-log data, together with petrophysical interpretation from 14 wells, are used in this competition. Nine training wells are numbered from 0 to 8: 15\_9-19, 15\_9-F-1, 15\_9-F-1A, 15\_9-F-1B, 15\_9-F-1C, 15\_9-F-4, 15\_9-F-5, 15\_9-F-10, 15\_9-F-11A. Five testing wells are numbered from 100 to 104: 15\_9-F-11B, 15\_9-F-11T2, 15\_9-F-12, 15\_9-F-14, 15\_9-F-15.

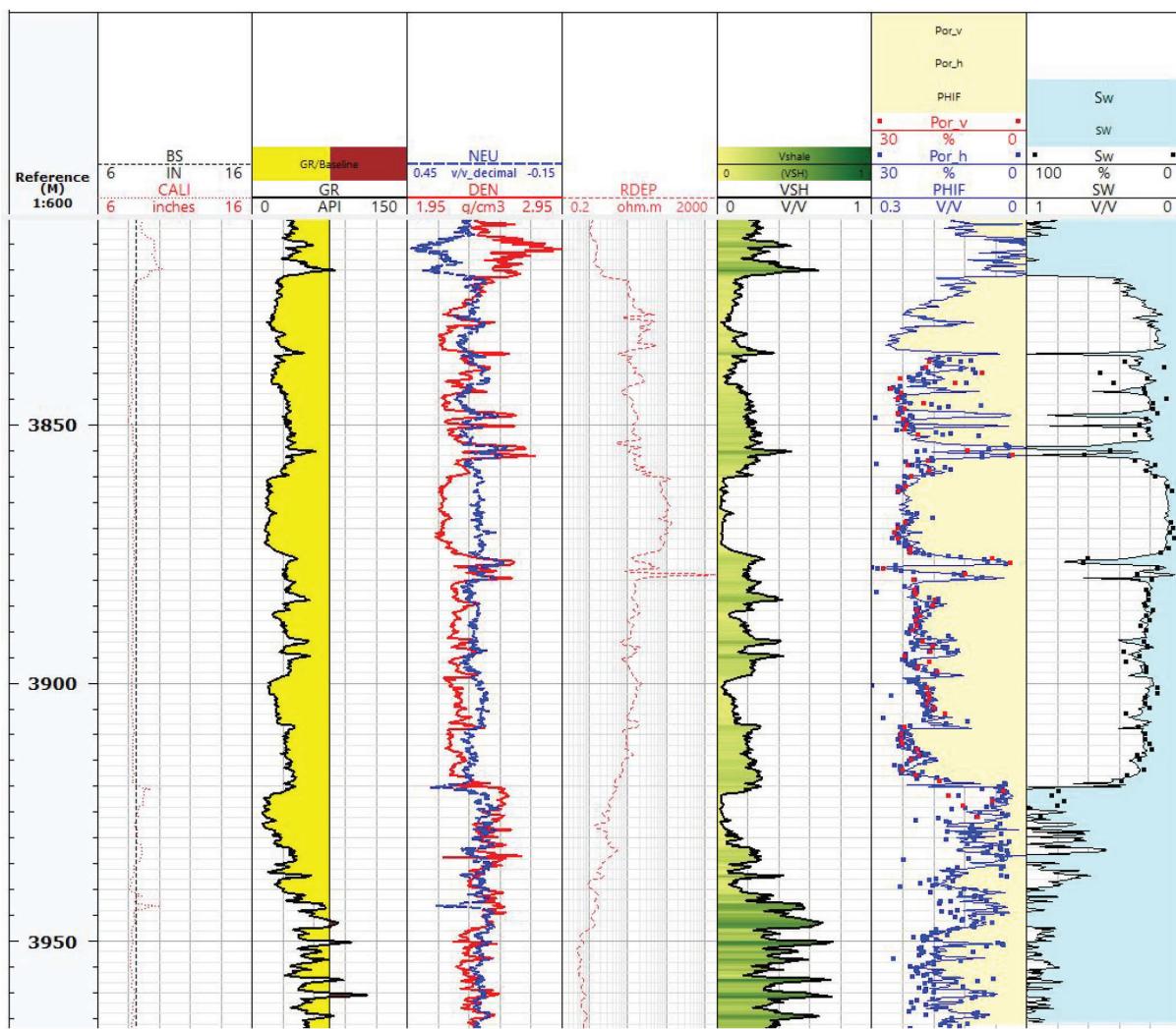
The location of the Volve Field is shown in Fig. 1, and the wells we selected from the Volve Field are shown as a structure map in Fig. 2. Figure 3 shows an example of cored well logs calibrated by the petrophysicists.



**Fig. 1**—Location of the *tutorial* oil field.



**Fig. 2—**Structural map with well locations (modified from Soljell and Lehne, 2006).



**Fig. 3—**Well logs and standard petrophysical interpretation in Well 15\_9-19.

## DATA PREPARATION AND PREPROCESSING

Following standard petrophysical workflow, the organization committee first assembled the well-log data. Quality control measures were taken to identify and address any issues with the data. Basic displays of individual logs and overlay plots of multiple curves were generated to start qualitative visualization of the reservoirs and formation types. As a next step, statistical analysis and plots were produced to quantify data attributes, as shown in the published tutorial in the *SPWLA Today* newsletter (Fu et al., 2021).

This exploratory data analysis provided crucial insights into data distributions, correlations, outliers, and other factors that guided subsequent decisions on data preprocessing and feature engineering. The visualizations and statistics revealed crucial details about inter-relationships between features and targets, variability between wells, and potential data quality issues. This information informed key steps like log selection, handling of missing values, and identification of anomalous data points. In this way, following standard workflows for initial data investigation and quantification provided the necessary foundation on which to build the machine-learning modeling approach.

In the process, we first visualize and perform exploratory data analysis on all available data. A total of 318,967 samples (data vectors) are loaded, and each of them consists of 17 columns. The description and unit of each column in the log are listed in Table 1.

### Feature Selection

Following the example in the tutorial, five commonly used logs (DEN, GR, NEU, PEF, RDEP) are selected together with target columns (PHIF, SW, VSH).

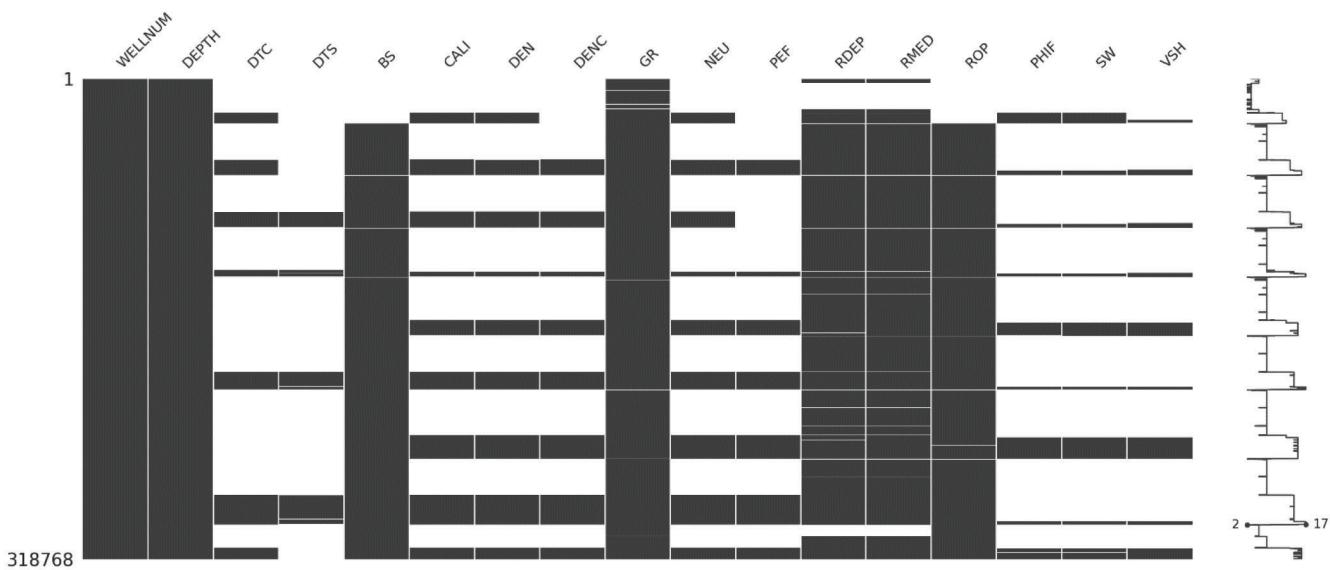
### Handling the Missing Data

Due to cost or operational difficulties, well logs are often missing at certain depths, as shown in Fig. 4. The value “–9999” is shown as the missing value in all features. One simple way to handle the missing values is removing all the data vectors that contain the missing values “–9999.” After removing all the missing values in the five commonly used logs (DEN, GR, NEU, PEF, RDEP) as inputs and outputs (PHIF, SW, VSH), there are 40,192 data vectors left, as shown in Table 2.

Note here we dropped the whole row if any column of the data point contains a missing value. As a result, only less than 12% of data are left. However, if you select different features as inputs or fill in the missing values, a lot more data can be utilized to build the machine-learning model.

**Table 1**—Column Names, Meanings, and Units of the Well Logs

Name	Description	Unit
WELNUM	well number	
DEPTH	depth	feet
DTC	compressional traveltime	microsecond per foot
DTS	shear traveltime	microsecond per foot
BS	bit size	inch
BVW	bulk volume of water	v/v
CAL	caliper size	inch
DEN	density	gram per cubic centimeter
GR	gamma ray	API
NEU	neutron	dec
PEF	photoelectric factor	Barn
RDEP	deep resistivity	Ohm meter
RMED	medium resistivity	Ohm meter
ROP	rate of penetration	meters per hour
PHIF	porosity	v/v
SW	water saturation	v/v
VSH	shale volume	v/v



**Fig. 4**—Missing values in features and target values.

Downloaded from <http://onpetro.org/petrophysics/article-pdf/65/01/1083391105/spwia-2024-v65n1a6.pdf/1> by Saudi Aramco user on 18 October 2024

**Table 2**—Statistical Attributes of the Training Data After Removing the Missing Values in Five Common Logs and Target Outputs

	<b>count</b>	<b>mean</b>	<b>std</b>	<b>min</b>	<b>25%</b>	<b>50%</b>	<b>75%</b>	<b>max</b>
<b>DEPTH</b>	40192	12131.01	1557.91	8494	11154.2	11837.27	13115.81	15566.93
<b>DTC</b>	17431	76.39	12.32	48.93	68.28	74.65	83.37	123.25
<b>DTS</b>	7198	131.02	14.55	74.82	123.15	131.9	138.54	193.84
<b>BS</b>	40192	8.5	0	8.5	8.5	8.5	8.5	8.5
<b>CALI</b>	40192	8.62	0.09	8.3	8.56	8.62	8.67	9.18
<b>DEN</b>	40192	2.41	0.16	1.63	2.26	2.42	2.55	3.09
<b>DENC</b>	40192	0.05	0.02	-0.49	0.03	0.05	0.06	0.26
<b>GR</b>	40192	39.13	24.11	4.59	19.38	34.89	51.05	129.34
<b>NEU</b>	40192	0.18	0.07	0	0.14	0.18	0.22	0.58
<b>PEF</b>	40192	5.18	1.72	-0.01	4.58	5.41	6.1	13.36
<b>RDEP</b>	40192	26.18	450.55	0.13	1.58	3.01	9.46	80266.8
<b>RMED</b>	40192	786.81	6788.72	0.14	1.65	3.54	11.56	62290.8
<b>ROP</b>	40095	23.43	7.76	0.34	18.82	24.98	29.91	46.9
<b>PHIF</b>	40192	0.14	0.08	0	0.07	0.14	0.22	0.4
<b>SW</b>	40192	0.66	0.36	0.01	0.29	0.78	1	1
<b>VSH</b>	40192	0.27	0.2	0	0.11	0.23	0.35	1

## Removing the Outliers

One of the findings from Table 2 is that the maximum values of all features are dramatically larger than their mean values, which indicates anomalies and outliers exist in the data set. It is helpful to remove those abnormal data points to improve the performance of the machine-learning models, especially for certain types of models like neural networks. In the tutorial (Fu et al., 2021), the Isolation Forest algorithm was used to remove the outliers in the training data. The Isolation Forest “isolates” observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. We used default values provided in the scikit learn

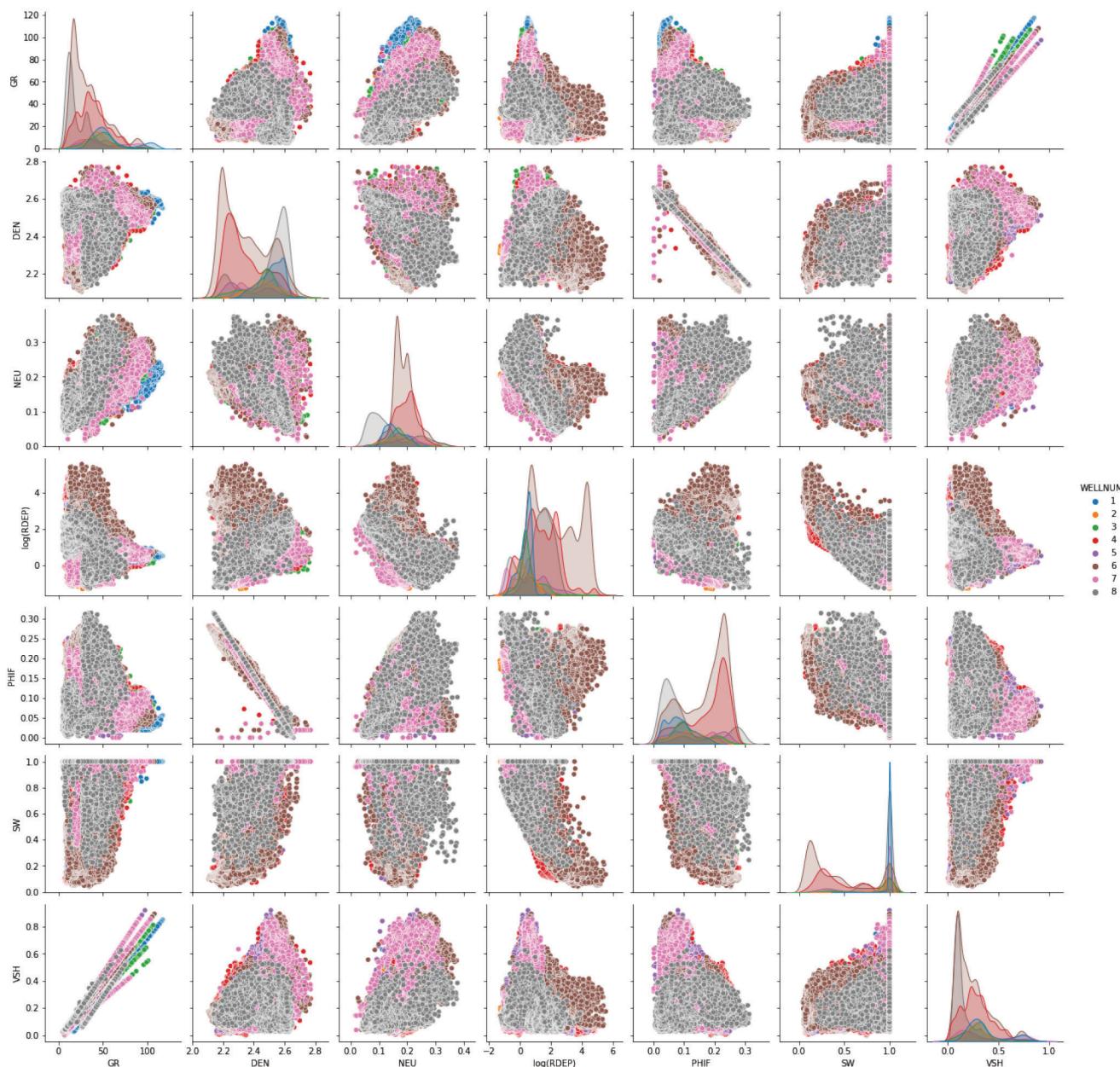
package. The number of base estimators in the ensemble is 100. The number of samples to draw from X to train each base estimator is 256. The amount of contamination of the data set, i.e., the proportion of outliers in the data set, is calculated and determined as in the original paper (Liu et al., 2008). Quality control of the log data is a crucial step to ensure good performance of the machine-learning model. More information on this can be found in Misra et al. (2019).

## Correlation

Crossplots are useful visualization tools to exhibit the correlations between features and targets. Figure 5 shows the crossplots between input features (gamma ray, bulk density,

neutron porosity, and logarithm of deep resistivity) and targets (porosity, water saturation, and shale volume). We use different colors to represent data from different wells. Certain well logs are highly correlated with the targets. For example, density (DEN) and porosity (PHIF) have a strong linear relationship, most likely using the density porosity equation but varying the grain density and fluid density. Another example is gamma ray (GR) versus shale volume (VSH), although the slopes of different wells are slightly

different, again seeming to use one of the myriads of shale volume from gamma ray equations. Another important observation is that the data distributions of each feature or target are quite different from well to well shown in crossplots in the diagonal positions. This indicates that well logs from different locations measure different geological structures or lithology. As a result, it's important to select data from wells or depths that are most relevant to the target values we are trying to predict.



**Fig. 5**—Plot pairwise relationships of the data set.

## MACHINE-LEARNING MODELS

All machine-learning models are available at <https://github.com/pddasig/Machine-Learning-Competition-2021>. Winning teams use different preprocessing methods and machine-learning models, which are summarized together with their corresponding RMSE scores in Table 3. Teams MoLPhy, AtwahAnalytics, and Jaehyuk\_Lee use ensemble models, which integrate different machine-learning models to yield more robust results. Team UTFE uses a different preprocessing method to impose stationary constraints for ML-based well-log interpretation (Pan, 2022). Team Tomsk uses the classical neural network method for petrophysical interpretation. Details about each team's work are summarized in this section.

### UTFE

Different wells have different borehole environments and see different formation properties. The limited resolution of logging tools and different environments, such as logging tools, vintage, vendors, and drilling fluid, can cause systematic bias in well-log measurements, resulting in non-uniqueness in well-log interpretation. To mitigate the error introduced by logging environments and decrease the non-uniqueness of interpretation, Pan et al. (2023) proposed the type well adaption method.

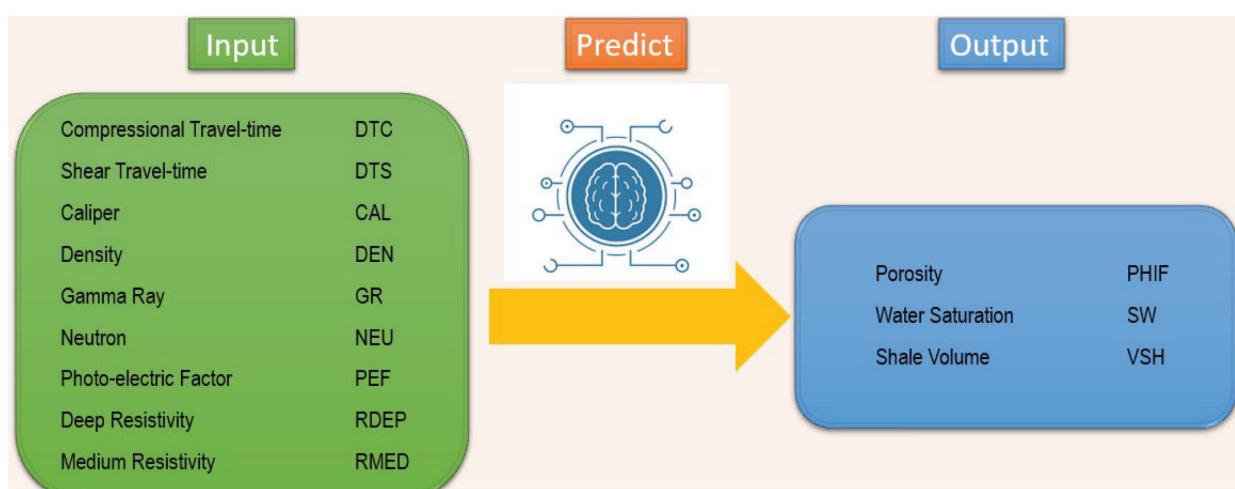
Instead of using well logs from all training wells as training data, which inevitably propagates the error introduced by different logging environments to the interpretation process, Pan (2022) suggested only using training wells that have similar formation properties and

borehole environment as the test well. The relationship between well logs and formation properties is expected to be similar in the test and selected training wells.

These training wells that adapt to test wells are identified as type wells. However, manually identifying the type well is time consuming and subject to the interpreter's experience. To mitigate the efforts for type well selection, they propose to use statistical distance and normalized dynamic time warping (DTW) distance to quantify the similarity between test wells and training wells. For wells with similar formations, type wells are selected as the well that has the lowest KL divergence (Pan, 2022; Pan et al., 2023) to a test well, e.g., saturation estimation in this competition, while for wells that cover different intervals of formations, DTW-based distance is used to find type wells that contain intervals similar to the test well, e.g., VSH prediction in this competition.

With the training data adapted to the test data, they build a simpler machine-learning model to perform the mapping from well logs to formation properties. In this competition, they found the relationship between well logs and formation properties is quite simple; therefore, they mainly used the K-nearest neighbors method for its simplicity, i.e., it only has one hyperparameter. Future work could be to test other types of machine-learning models.

They used a dummy semi-supervised learning method to improve the interpretation of intervals with missing data, i.e., use adjacent, interpreted intervals as training data to make predictions. More advanced methods, such as long short-term memory (LSTM) and other statistical methods, can be used to improve the results.



**Fig. 6**—A diagram shows the general workflow to predict reservoir properties based on well logs using a machine-learning approach.

**Table 3**—Methods and Scores From the Top Five Teams

Rank	Team Name	Methods	RMSE
1	UTFE	Linear regression with well selections	0.0525
2	MoLPhy	Ensemble model	0.0602
3	Tomsk	Neural network	0.0634
4	AtwahAnalytics	ExtraTree and Catboost regression	0.0667
5	Jaehyuk_Lee	LightGBM and XGBoost	0.0775

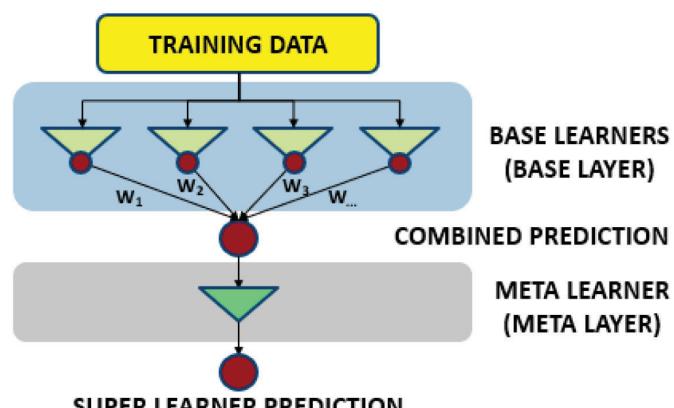
To summarize, this team first finds a linear correlation between GR and VSH. However, the coefficients of linear models vary among wells. To mitigate the prediction error introduced by nonstationarity, this team finds type well and test well pairs that minimize divergence and train different linear models for each pair. The DTW method is used to perform zonation, which decreases the nonstationary error along a single well. The same type of well selection method is used for water saturation estimation, but the linear model is replaced with a KNN (K-nearest neighbor) model. The semi-supervised learning method is used for data imputation, and only the local data from the same well is used for data imputation to avoid error-introduced inter-well nonstationarity. The porosity is estimated without preprocessing because this team found the error for porosity prediction is low compared to the other two petrophysical properties.

### MoLPhy

Team MoLPhy found that managing nonrepresentative/exclusive lithology plays a crucial step in data preparation. Clustering of data to achieve better prediction involving and merging both train and test data sets was tried, but group-specific modeling did not bring lower RMSE after multiple submissions. So, after getting discouraging results from clustering, it was realized (based on plotted logs from train/test wells) that test wells contained special intervals, most likely representing source rock with abundant organic matter, that is not well represented in training wells mainly due to the missing interpretation. Over such intervals, unlike VSH and PHIF (only in shales), SW prediction is not acceptable, so basically, with simple flagging (based on lithology indicators cutoffs), they selected such intervals (naming as Cluster-1) on the way to mainly have coverage of those source rock intervals in test wells, and also one can identify them in training wells, too, in order to ensure feasible training. That means running the training process in two main steps: firstly, over the entire interval fitting to

all parameters, and secondly, over only Cluster-1, and only targeting SW because that is the affected parameter (the other and more problematic one is VSH, which, in our opinion, brings the majority of overall RMSE after analyzing train/validation errors)

They used a stacking ensemble model, Super Learner (Van der Laan et al., 2007), from mlens library (Flennerhag, 2017) that comprises multiple regressors with base learners and meta-model in dedicated layers. The structure of the machine-learning model is shown in Fig. 7. The selection of base models, meta learner, and their parameters were probed by the final RMSE. The final model includes LinearRegression, ElasticNet, DecisionTreeRegressor, KNeighborsRegressor, BaggingRegressor, ExtraTreesRegressor, XGBRegressor, and LGBMRegressor as base learners and ExtraTreesRegressor as a meta learner. The advantage of such a stacking method is that by weighting each model prediction as per their performance (i.e., regression score) and reducing the bias of each model, a more robust (super)model can be achieved by giving the lowest possible prediction error with lower risk of overfitting supported by internal k-fold cross validation on the training subset.

**Fig. 7**—Structure of Super Learner model.

In their approach, the prediction of target parameters is made sequentially (not only because Super Learner can only handle 1D target array), unlike simultaneous solutions. It means that in the order of petrophysical workflows (so not in charge of provided data), it fits VSH the first time, then appends VSH to the input predictor data frame and fits PHIF, and finally, SW. Similarly, the prediction phase predicts VSH, appends it to the input data, then predicts PHIF, and finally, SW. Of course, that brings prediction error into the system, but they believe that it gets much more information with sequential parameter re-usage, which aids in fit and prediction. Integration of already predicted parameters into the training set (for subsequent training) gives higher data coverage, bringing in more information due to the connection between target parameters. The other aspect of such a sequential solution is that users can specify a predictor's list specific to the actual target parameter fit in each step. It is not necessarily true that all predictors are used for all targets, as not all targets depend on all predictors in the same way, e.g., GR log is more important in the prediction of VSH than SW.

They also applied data transformation to the testing data. Inspection of train and test predictors/features distribution indicated mismatches between them, hence causing a model regularization problem, so that with the help of a simple histogram correlation optimization function (only applying linear transformation with a multiplier and shift parameter), the test input data distribution was transformed to be similar to train data. That further lowered the final RMSE.

The main takeaway from their experience is that data matter more than any model architecture. Even a highly hyperparameter-tuned model can give only low error reduction compared to the improvement effect of optimized and harmonized data. Insufficient quantity or non-representativeness of training data (sampling bias, model generalization issue), poor quality (erratic values, outliers, noise), and incompleteness (data gaps) are all reasons for a much higher impact on final ML model performance.

### **Tomsk**

The preprocessing applied to the raw data includes taking the logarithm of the resistivity logs to linearize them, missing values and outliers handling, and scaling. All missing values outside the targets were dropped. They

applied a backward fill algorithm to fill the remaining missing values in features. As a method for anomalies detection, they selected the Isolation Forest algorithm. Identified anomalies were dropped from the training data.

They applied two scalers for the features: RobustScaler, a method that scales features using statistics that are robust to outliers to scale the data according to the 0.02 and 0.98 quantiles and then transform features by scaling each feature to a range determined by the minimum and maximum values of the training data set. Additionally, they also cut off target values  $> 1$  in the training data set, as they have noticed that some samples within targets are not in range from 0 to 1, which is physically impossible. The MinMaxScaler was applied to all targets.

From the exploratory data analysis (EDA), a range of dependencies was observed, which were applied for the post-prediction correction. The DEN, GR, and the logarithm of RDEP (RDEP\_log) distribution show that the density log for test wells can have values up to 3.2, while the upper limit for the training wells is about 3.1. All the samples within training data with  $\text{DEN} > 3$  correspond to the porosity value equal to 0.02. Therefore, they expect the same porosity for samples with  $\text{DEN} > 3$  within the test data.

Values of GR for the train and validation data set vary significantly. The maximum for training wells is 129.3 and for validation wells is 1711.4. In a pair plot GR vs. VSH, the linear correlation between GR and VSH exists, so the high GR values are associated with high clay content. They assume that samples in the validation data set with a GR higher than the maximum GR within training data should be equal to 1.

RDEP\_log values of the validation data set can reach values less than -2. The low values of RDEP\_log indicate high values of SW. It is also confirmed that there is a high portion of samples in the train data with low resistivity, where SW is equal to 1. Plotting the mean values of RDEP\_sqrt and mean values of SW for each well, they define two groups of wells. The first group is characterized by low resistivity and high SW (Wells 1, 7, 2, 0, 3, 8); the second group, on the contrary, is characterized by high resistivity and low SW (Wells 4, 5, 6). Plotting only the mean RDEP\_sqrt for validation wells shows that Well 100 (15\_9-F-11T2) corresponds to the first group. Therefore, they assume that the whole interval of Well 15\_9-F-11T2 is equal to 1.

They tried two clustering approaches in order to improve the metric but with limited or no success:

1. Hierarchical clustering
2. They applied the RANSAC (Random Sample Consensus) algorithm in order to delineate samples that correspond to lines with a specific slope and then applied hierarchical clustering of mean well-log values in order to obtain clusters that correspond to specific classes. They had a hypothesis that a classifier could be built to predict the same classes on test data and then use the class as a feature or build separate models for each class, but the classifier did not work well.
3. Binary classification of SW = 1 and all the others

They were inspired by the results of <https://towardsdatascience.com/zero-inflated-regression-c7dfc656d8af> and predicted samples with SW = 1 using the CatBoost classifier, and then they replaced the value with 1 in samples that the classifier predicted as equal to 1. Despite the quite good accuracy of the classifier, this approach does not improve the metric of the competition.

*Feature Engineering.* Only one additional feature was generated in addition to the set of CALI, GR, DEN, and RDEP\_log curves. Based on the DEN log, they discretized the training data into 10 bins with a fixed number of points. Within the resulting bins, they calculated the mean porosity, and using the mean porosity within each bin, they reconstructed this curve for the validation data.

To build the machine-learning model, the FFNN with two hidden layers and three outputs were used. The model architecture was chosen following the idea of linear interdependences between target and feature well logs, although they conducted a few experiments with a higher number of hidden layers. The output layer consisted of three targets. Sequential optimization of neural networks with a single output is something they see as worth trying further. The hyperparameters of the model were obtained using an automated hyperparameter search tool named Optuna with some additional manual tuning. The best RMSE score (without application of post-processing functions) obtained after the model performance is 0.069, with small fluctuations depending on the environment. The optimization process took around 10 hours of wall-clock time using a Tesla GPU on a DGX workstation.

*Post-Processing.* The primary model prediction result was processed by the application of several corrections stated in the EDA section. The porosity for the samples characterized with DEN values higher than 3 was filled with 0.02. Additionally, the porosity is clipped between 0 and 0.95 percentiles.

The water saturation in Well 15\_9-F-11T2 is filled with 1, as it has high values of RDEP\_log compared to other test wells. The SW in other wells was corrected by the summation with the error received from the train wells prediction. Furthermore, samples that have RDEP\_log values less than -2 are also filled with 1. For the volume of shale, only one correction was applied. If the samples with GR value exceed the max GR in the train data set, its VSH is replaced with 1.

After the post-process correction, the score is improved from 0.069 to 0.0634 for the validation set.

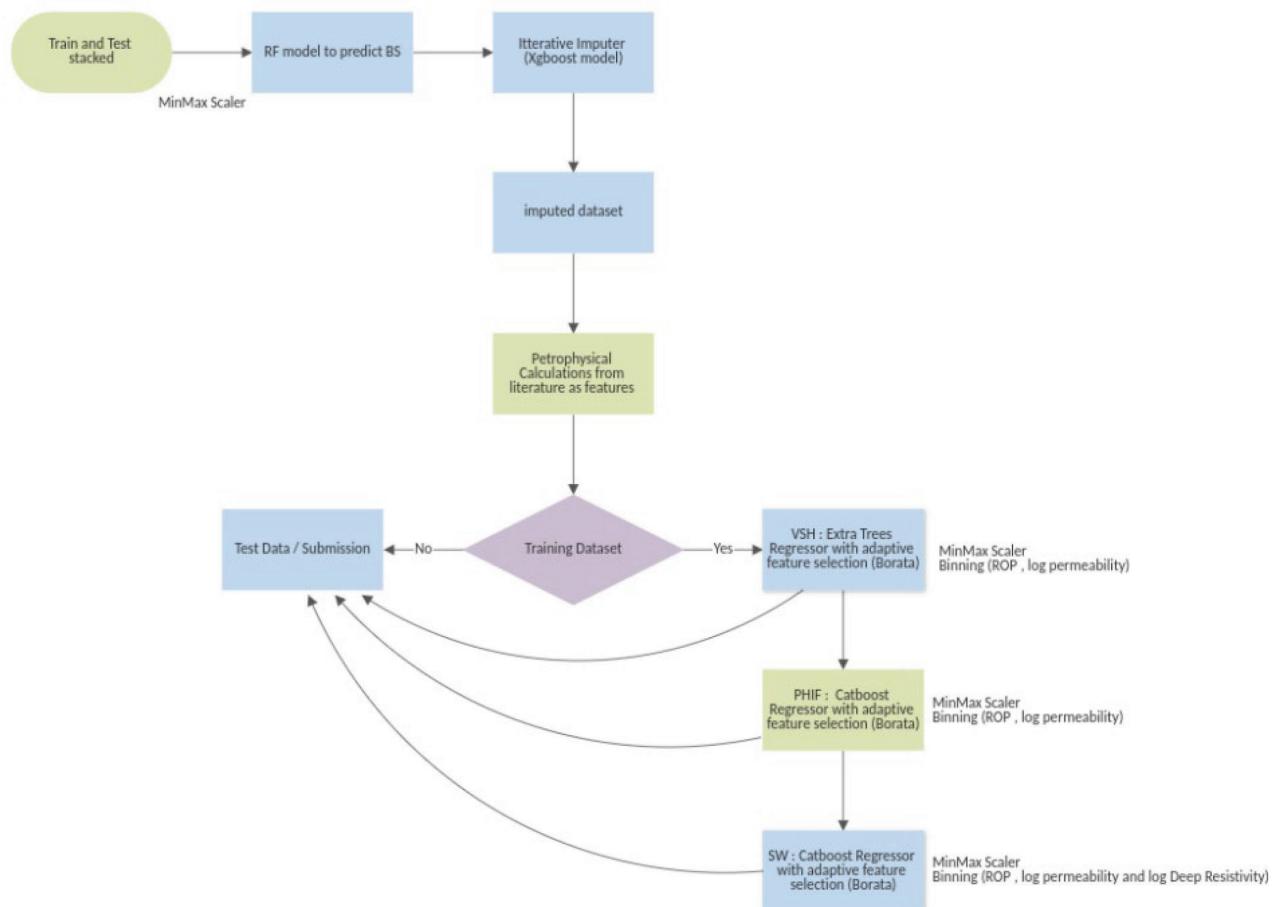
### AtwahAnalytics

When developing a predictive model using machine learning or statistical modeling, feature engineering refers to the process of leveraging domain expertise to choose and convert the most important variables from raw data. In their solution, they applied the following equations to use domain knowledge for feature engineering. Most of the functions used are retrieved from the following GitHub link:  
<https://github.com/whamlyn/auralib/blob/01d64e25018fa249b3f901700428e9cb211d803cauralib/pp.py>

1. The first step is to build extremely randomized trees, an ensemble supervised machine-learning method, to predict VSH while dynamically choosing the best performing features using a feature selection method, "boruta." Some additional processing was done on the input features, such as rate of penetration (ROP) and log-transformed average permeability, which is known as numerical feature binning.
2. The second step is to build using a machine-learning method based on gradient boosting on decision trees, CatBoost Regressor, to predict PHIF following the same procedures as described in Step 1.
3. The third step is to build a CatBoost Regressor to predict SW following the same procedures as described in Step 1.

The workflow is summarized in Fig. 8:

## Modeling Flow Chart



**Fig. 8**—The workflow of Team AtwahAnalytics.

### Team Jaehyuk\_Lee

This team found this data set had a substantial portion of missing data, and most of the features have many outliers, which can distort the true distribution of data. Before training a model, they focused on understanding each feature of the data set and removing all missing data to help the efficiency of modeling time.

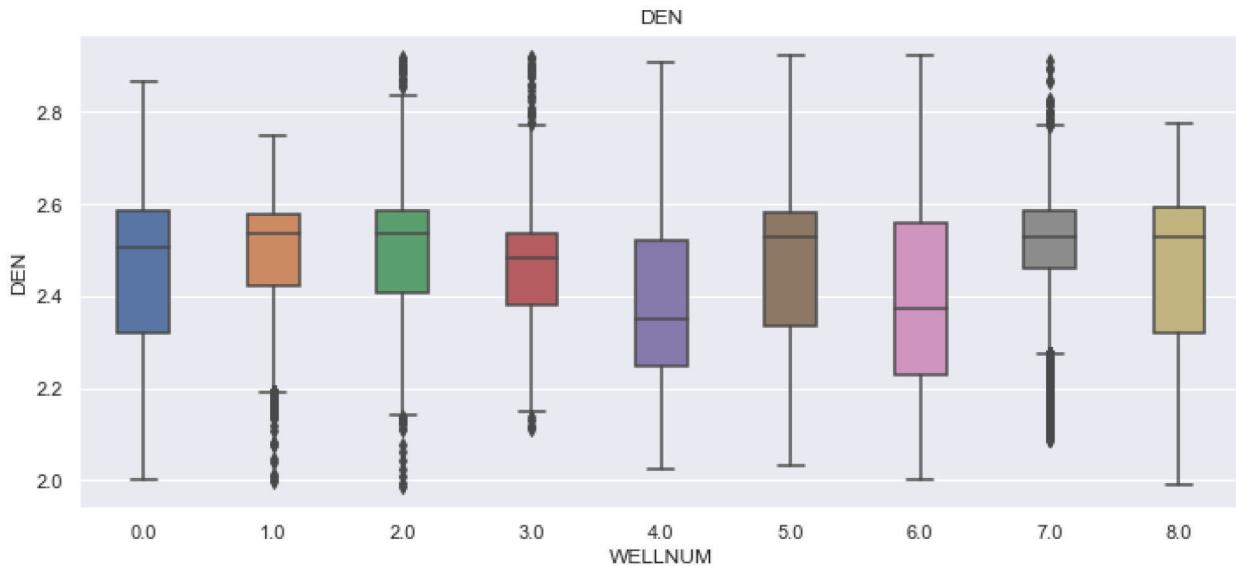
First, they checked the overview of the data set. The results showed a lot of -9999 numbers in the data set. It means a large amount of missing data in the data set and should be handled properly. After these missing values were replaced with not a number (NaNs), they visualized them to see the missing values (Fig. 4).

They found the target columns (PHIF, SW, VSH) had more than 85% of the data that were missing, and also, other columns, such as DEN, NEU, PEF, DEN, and DENC, had

more than 70% of the data missing. They should be removed or imputed to fill in valid data.

For outlier detection, they checked each column's data distribution using a boxplot. Each feature (column) has a considerable size of spread of distribution of data. One example is shown in Fig. 9. These large values were assigned as an outlier when it has  $\pm 3$  standard deviation from their mean value. Then, they were removed from the data set and will not be calculated by the machine-learning model.

In the data-handling process, they parameterized all possible combinations, such as the number of columns, the degree of outlier detection criteria, and missing value imputation methods, and then applied that to the training process. These parameters were optimized based on model performance after training and prediction.



**Fig. 9**—Boxplot for density shows the data distribution.

As referencing the missing values only in target columns, 87% of rows of the data set were removed from the original data set. Nevertheless, “DTC” and “DTS” still have a substantial portion of missing data (DTC: 54.43%, DTS: 82.97%), then these two columns are dropped. Although two columns were removed, NEU data can represent them because NEU data are highly correlated with DTC and DTS data based on correlation analysis.

For model training, they tried various modeling algorithms, such as linear regression, SVM, and tree-based algorithm, and a few deep-learning frameworks, such as Keras and PyTorch. They narrowed down these algorithms comparing with train-test score and their prediction rankings.

Below is their training procedure:

1. Data preparation: splitting data set into training and test sets (80/20)
2. Model selection: using gradient boosting framework that uses tree-based learning algorithm LightGBM regressor, eXtreme Gradient Boosting (XGBoost) regressor, Random Forest regressor, Linear regressor, Linear Support Vector regressor.
3. Cross validation: cross validation with five splits
4. Hyperparameter tuning: optimizing the hyperparameter of the model using a cross-validated grid search over a parameter grid

After comparing their results in the initial stage, some tree-based models showed better results, including LightGBM and XGBoost. They optimized the hyperparameters of the XGBoost model and LightGBM model using GridSearchCV.

## RESULTS COMPARISON

Table 3 shows the RMSE achieved by the top five teams in the contest. All teams have outperformed the benchmark set by the committee.

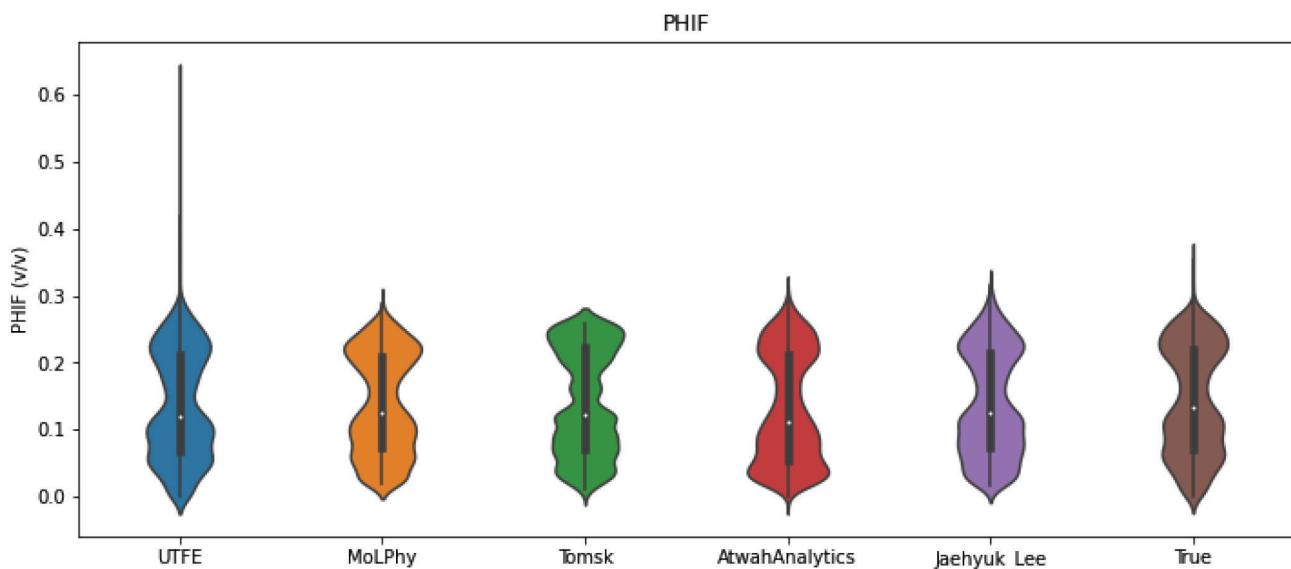
We plotted the predicted results in the test well submitted by the top five teams. Both a crossplot and violin plot, a hybrid of a boxplot and a kernel density plot, are shown for a better view and comparison. Different colors are used to differentiate the teams.

Figures 10 to 12 are the violin plots of the predicted porosity, water saturation, shale volume, and their true values. Overall, the top five teams provide accurate predictions, which have a similar distribution as the true ones. A violin plot is a combination of boxplot and kernel density estimate. The violin plot features a kernel density estimation of the underlying distribution.

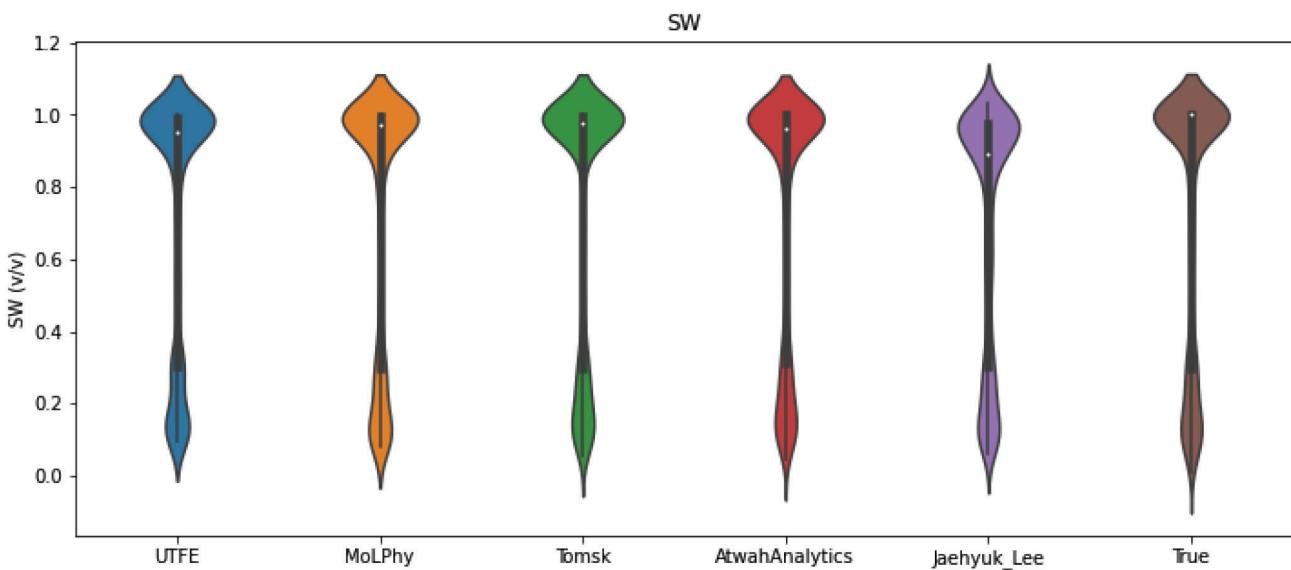
Figures 13 to 15 are the crossplots of the predicted porosity, water saturation, shale volume, and their true values. The x-axis represents the true value, while the y-axis is the predicted value. The top left panel shows that the predicted

values of the top five solutions together overlap with each other. The other five panels display the predictions from the team from first place to fifth place separately. The black line is  $y = x$ , which represents the case when the predictions match with true values perfectly. We can see that most of the data points are reasonably distributed along the  $y = x$  line. Although predictions from all the teams generally match well with the true values, there are still certain predictions that are underestimated or overestimated. The discrepancy is obvious for large values, which show as straight lines.

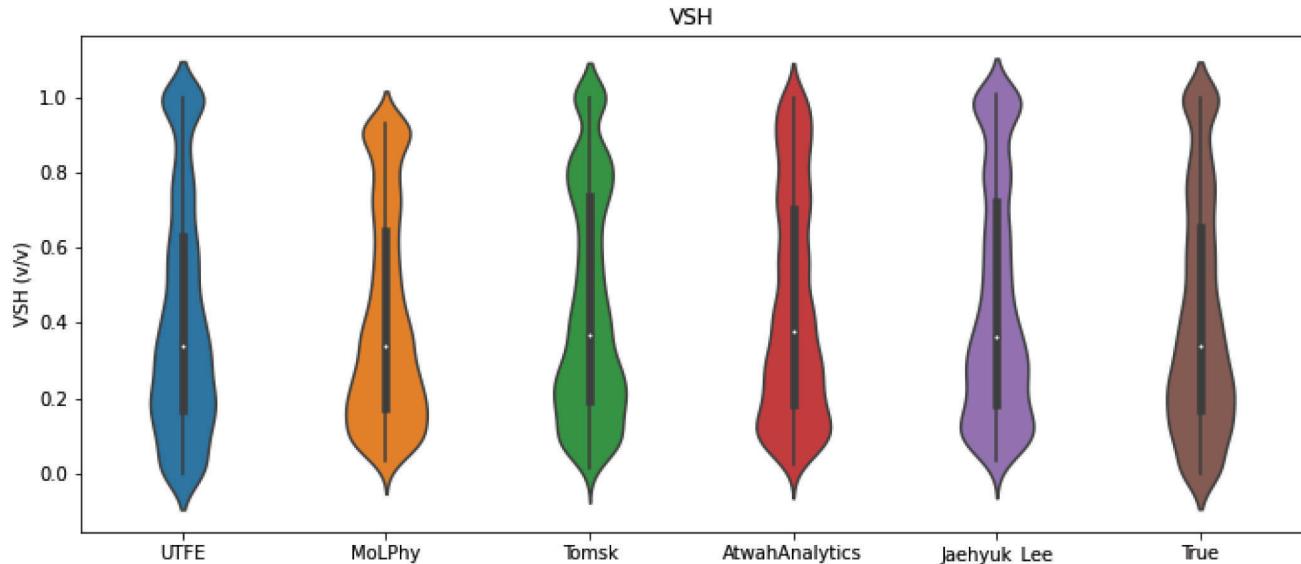
Figures 16 to 18 show the comparison of results in a log profile over a deep interval as an example. The true values represent the petrophysical interpretations of VSH, PHIF, and SW from experts based on the same criteria used in the training data. Various solutions provide different estimations with both underestimation and overestimation at different depths. In Fig. 18, we notice that the predicted shale volume in the deep part is systematically lower than the measured values, which indicates the machine-learning models are still not complex enough to capture maximum variabilities.



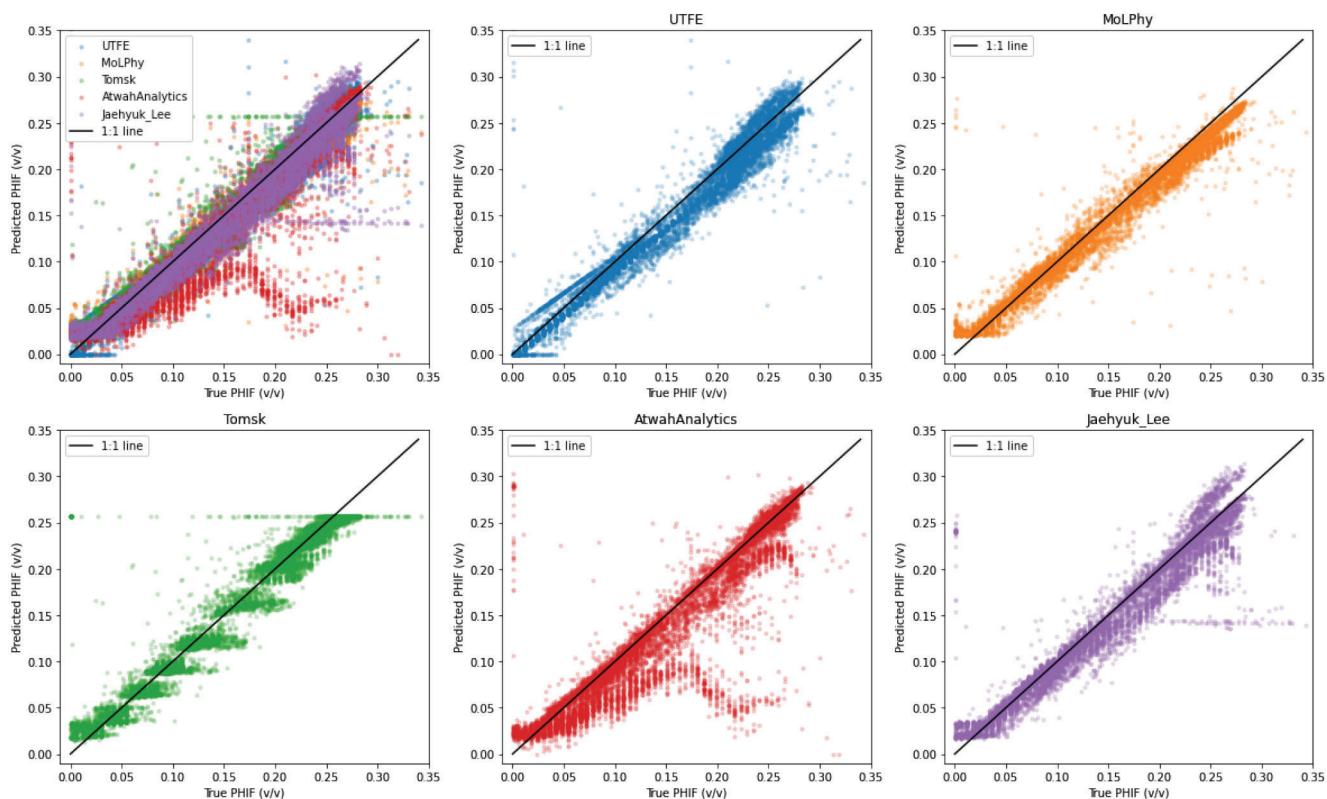
**Fig. 10**—Distribution of the predicted and true porosity.



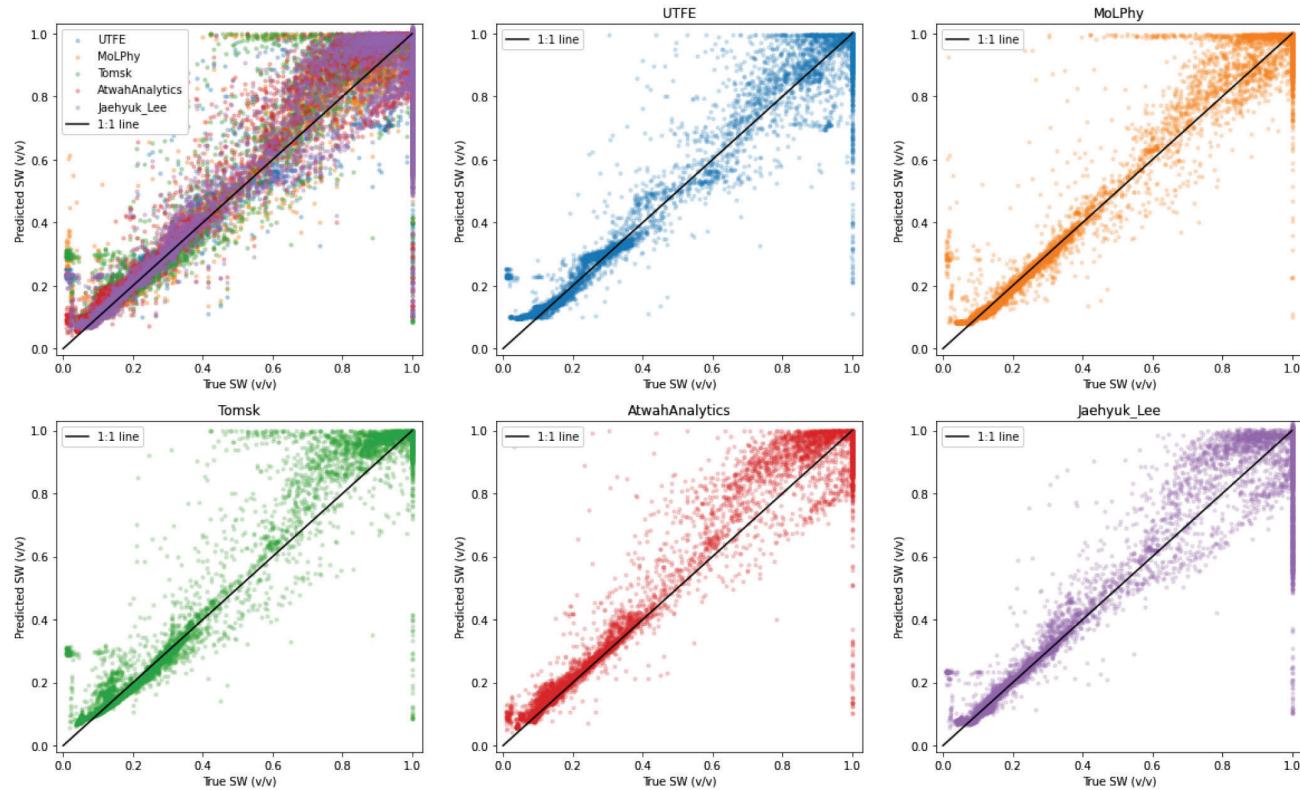
**Fig. 11**—Distribution of the predicted and true water saturation.



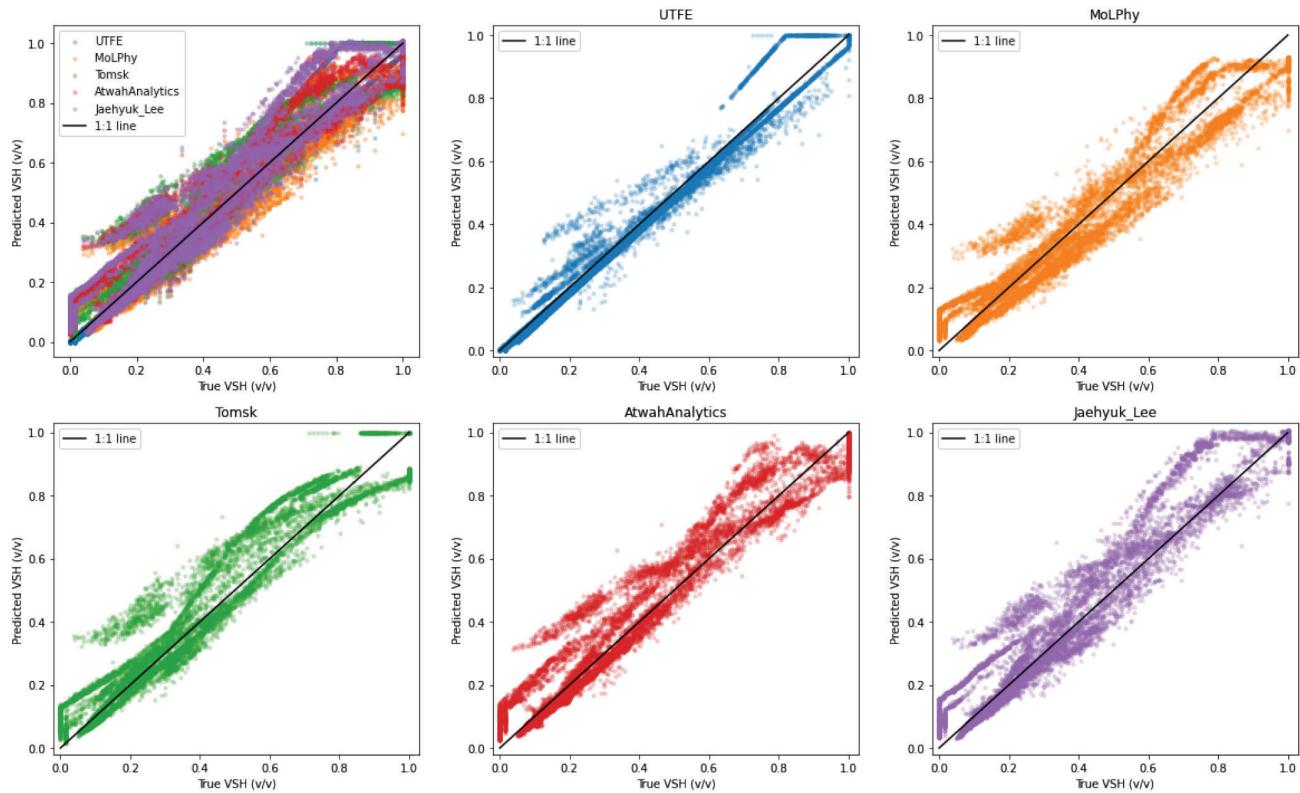
**Fig. 12**—Distribution of the predicted and true shale volume.



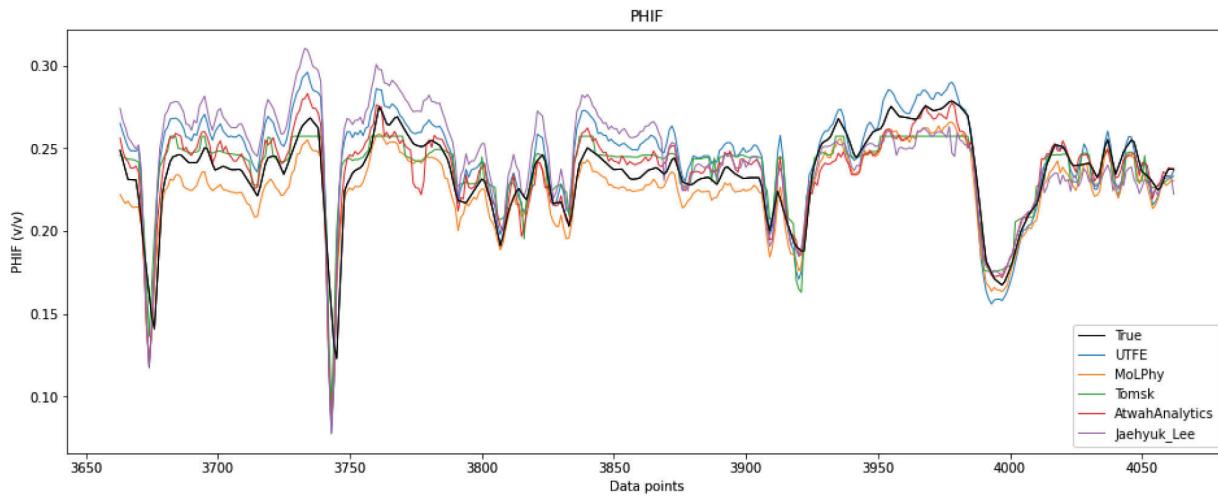
**Fig. 13**—Crossplots of the true and predicted porosity.



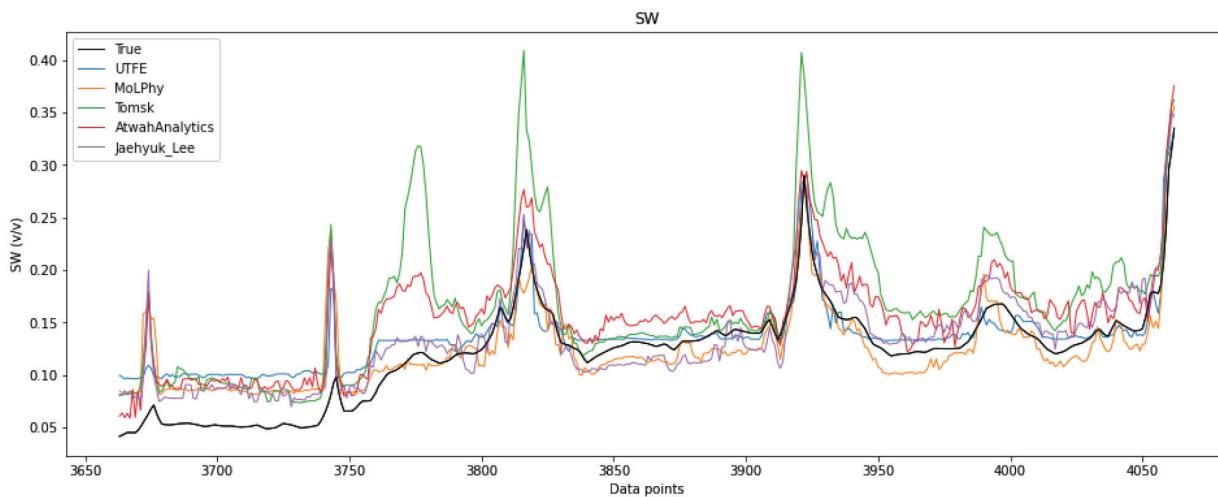
**Fig. 14**—Crossplots of the true and predicted water saturation.



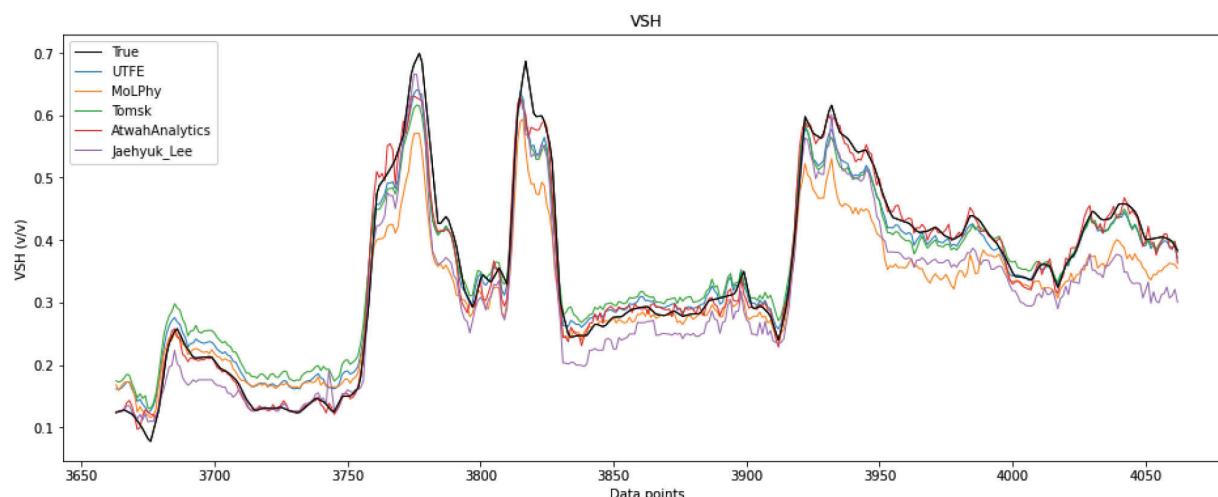
**Fig. 15**—Crossplots of the true and predicted shale volume.



**Fig. 16**—Comparison of the predicted and true porosity in a deep interval.



**Fig. 17**—Comparison of the predicted and true water saturation in a deep interval.



**Fig. 18**—Comparison of the predicted and true shale volume in a deep interval.

## DISCUSSION

In this study, different teams were able to use various machine-learning workflows on a practical petrophysical problem: estimating reservoir properties using well logs from multiple wells from the same area. Techniques like data set preparation and quality assurance, feature engineering with outlier handling, and clustering are applied, followed by training and testing a regression model, and finally, blind testing (similar to the real-world deployment) of the model on the hidden data set.

As noted in the introduction, the subjectivity of the “ground truth” interpretation data used for training and evaluation provides inherent uncertainties. Given this limitation, we cannot make definitive claims that one model outperformed another based solely on the numerical scores. The rankings provide a general guideline but do not necessarily prove the superiority of one model over another. This point should be kept in mind when assessing the discrepancies between model results and the benchmark interpretation later in the discussion. Differences from the benchmark do not necessarily indicate the machine-learning results are inaccurate, considering the biases that may exist in the benchmark data itself.

Various models have been chosen and adopted by the different teams. The results suggest that the choice of the machine-learning model itself might not be the key to the success of predicting on the new data set. Rather, attention should be focused on other methods to improve the performance and stability of the model. These areas should include but not be limited to only selecting training wells that are analogous to the test well, appropriate handling of outliers and anomalies, exploring multiple different models for data from different wells, focus input data quality control, and data preparation.

## SUMMARY

In this contest, all participating teams demonstrated using a machine-learning workflow on a practical petrophysical problem: preparing a data set, training and testing a regression model, and finally, testing the model on unseen data. Libraries and open-source tools, such as scikit-learn, provide powerful algorithms that can be applied to problems with a few lines of code, which greatly helps

to facilitate the research of data science in the petrophysics area. In addition to the procedures mentioned above, many other methods may be applied to improve the performance and stability of the model, such as applying better treatments to the missing values and anomalies, training different models for zones with different lithologies, training other regression models, and/or combining them.

For more details about the data and code, please check the GitHub repository: <https://github.com/pddasig/Machine-Learning-Competition-2021>.

## ACKNOWLEDGMENTS

A note of thanks goes to Equinor for releasing the Volve data set. We also thank the members of the SPWLA PDDA SIG ML Contest Committee for their contributions.

## NOMENCLATURE

### Abbreviations

DEN	= density
DTW	= dynamic time warping
EDA	= exploratory data analysis
GR	= gamma ray
KNN	= K-nearest neighbors
ML	= machine learning
PDDA	= Petrophysical Data-Driven Analytics
PHIF	= porosity
RANSAC	= Random Sample Consensus
RDEP	= deep resistivity
RDEP_log	= logarithm of RDEP
RMSE	= root-mean-squared error
SIG	= special interest group
SPWLA	= Society of Petrophysicists and Well Log Analysts
SW	= water saturation
VSH	= shale volume
XGBoost	= eXtreme Gradient Boosting

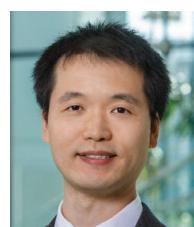
### Symbols

$i$	= index of the sample
$m$	= number of samples
$\hat{y}_i$	= predicted value
$y_i$	= true value

## REFERENCES

- Anemangely, M., Ramezanzadeh, A., Amiri, H., and Hoseinpour, S.A., 2019, Machine Learning Technique for the Prediction of Shear Wave Velocity Using Petrophysical Logs, *Journal of Petroleum Science and Engineering*, **174**, 306–327. DOI: 10.1016/j.petrol.2018.11.032.
- Deng, T., Ambía, J., and Torres-Verdín, C., 2020, Bayesian Method for Rapid Multi-Well Interpretation of Well Logs and Core Data in Unconventional Formations, Paper 5040, *Transactions, SPWLA 61st Annual Logging Symposium*, Virtual Online Webinar, 24 June–29 July. DOI: 10.30632/SPWLA-5040.
- Flennerhag, S., 2017, ML-Ensemble, URL: <http://ml-ensemble.com/>. Accessed November 14, 2023.
- Fu, L., Xu, C., Yu, Y., Ashby, M., McDonald, A., and Dai, B., 2021, Well-Log-Based Reservoir Property Estimation With Machine Learning: A Tutorial for the PDDA 2021 ML Contest, *SPWLA Today*, **4**(6), 42–49. URL: [https://www.spwla.org/Documents/Newsletter/SPWLA\\_Today\\_Newsletter\\_Vol4\\_No6.pdf](https://www.spwla.org/Documents/Newsletter/SPWLA_Today_Newsletter_Vol4_No6.pdf). Accessed November 14, 2023.
- Khan, M.R., Tariq, Z., and Abdulraheem, A., 2018, Machine Learning Derived Correlation to Determine Water Saturation in Complex Lithologies, Paper SPE-192307 presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 23–26 April. DOI: 10.2118/192307-MS.
- Li, W., Fu, L., and AlTammar, M.J., 2022, Well Log Prediction Using Deep Sequence Learning, Paper IGS-2022-184 presented at the ARMA/DGS/SEG International Geomechanics Symposium, Abu Dhabi, UAE, 7–10 November. DOI: 10.56952/IGS-2022-184.
- Liu, F.T., Ting, K.M., and Zhou, Z.H., 2008, Isolation Forest, 2008 Eighth IEEE International Conference on Data Mining, 413–422, Pisa, Italy, 15–19 December. DOI: 10.1109/ICDM.2008.17.
- Luycox, M., Wheelock, B., Kadir, H., Oyewole, E.O., Ijasan, O., and McLendon, D., 2022, Core-Calibrated Multi-Mineral Interpretation in Reservoirs With Complex Mineralogy, *SEG Global Meeting Abstracts*, 2431–2445. DOI: 10.15530/urtec-2022-3721250.
- Misra, S., Osogba, O., and Powers, M., 2019, Unsupervised Outlier Detection Techniques for Well Logs and Geophysical Data, Chapter 1, in Misra, S., Li, H., and He, J., authors, *Machine Learning for Subsurface Characterization*, Elsevier, 1–37. ISBN: 978-0-12-817736-5.
- Oh, S., Noh, K., Yoon, D., Seol, S.J., and Byun, J., 2019, Cooperative Deep Learning: Seismic-Constrained CSEM Inversion for Salt Delineation, *SEG Technical Program Expanded Abstracts*, 1055–1059. DOI: 10.1190/segam2019-3208029.1.
- Pan, W., 2022, Reservoir Description via Statistical and Machine-Learning Approaches, Doctoral dissertation, The University of Texas, Austin, Texas. DOI: 10.26153/tsw/44086. URL: <https://hdl.handle.net/2152/117205>. Accessed November 14, 2023.
- Pan, W., Torres-Verdín, C., Duncan, I.J., and Pyrcz, M.J., 2023, Improving Multiwell Petrophysical Interpretation From Well Logs via Machine Learning and Statistical Models, *Geophysics*, **88**(2), D159–D175. DOI: 10.1190/geo2022-0151.1.
- Singh, M., Makarychev, G., Mustapha, H., Voleti, D., Akkurt, R., Daghar, K.A., Mawlod, A.A., Al Marzouqi, K., Shehab, S., Maarouf, A., El Jundi, O., and Razouki, A., 2020, Machine Learning Assisted Petrophysical Logs Quality Control, Editing and Reconstruction, Paper SPE-202977 presented at the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, UAE, 9–12 November. DOI: 10.2118/202977-MS.
- Solfjell, E., and Lehne, K.A., 2006, Sleipner Øst and Volve Model 2006 Hugin and Skagerrak Formation Petrophysical Evaluation, URL: <https://www.equinor.com/energy/volve-data-sharing>. Accessed November 14, 2023.
- Van der Laan, M.J., Polley, E.C., and Hubbard, A.E., 2007, Super Learner, *Statistical Applications in Genetics and Molecular Biology*, **6**, article 25. DOI: 10.2202/1544-6115.1309.
- Xu, C., Misra, S., Srinivasan, P., and Ma, S., 2019, When Petrophysics Meets Big Data: What Can Machine Do?, Paper SPE-195068 presented at the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 18–21 March. DOI: 10.2118/195068-MS.
- Xu, C., Fu, L., Lin, T., Li, W., and Ma, S., 2022, Machine Learning in Petrophysics: Advantages and Limitations, *Artificial Intelligence in Geosciences*, **3**, 157–161. DOI: 10.1016/j.aiig.2022.11.004.
- Yu, Y., Xu, C., Misra, S., Li, W., Ashby, M., Pan, W., Deng, T., Jo, H., Santos, J.E., Fu, L., and Wang, C., 2021, Synthetic Sonic Log Generation With Machine Learning: A Contest Summary From Five Methods, *Petrophysics*, **62**(4), 393–406. DOI: 10.30632/PJV62N4-2021a4.
- Zhang, L., and Zhan, C., 2017, Machine Learning in Rock Facies Classification: An Application of XGBoost, *SEG Global Meeting Abstracts*, 1371–1374. DOI: 10.1190/IGC2017-351.

## ABOUT THE AUTHORS



**Lei Fu** is currently a data scientist at the Aramco Americas Houston Research Center, which he joined in 2021. His research is focused on developing machine-learning algorithms and applications for geoscience and engineering problems. Prior to joining Aramco, he was a research geophysicist and data scientist with SLB from 2019 to 2021. Lei received a PhD degree from Rice University in 2016 and an MS degree from the University of Utah in 2011. He currently serves as the Vice President of the SPWLA Petrophysical Data-Driven Analytics Special Interest Group (PDDA SIG).

He has served as the Chief Event Organizer and Treasurer of PDDA SIG (2021–2022) and led the SPWLA PDDA 2021 Machine-Learning Competition. Lei won second place in the 2020 SPWLA PDDA Machine-Learning Competition.



**Yanxiang Yu** is currently an applied scientist for Amazon Web Services. He previously has over 7 years of machine-learning experience in the energy industry, including working for companies like Shell and GOWell International. He has served as the chair for SPWLA PPDA SIG from 2020–2022 and advisor since 2022.



**Chicheng Xu** joined the Aramco Houston Research Center in 2017 and is working as a research petrophysicist in the AI Technology Group. His research focuses on petrophysics intelligence and automation using advanced computational techniques and data analytics for interpretation, classification, and modeling based on multiscale subsurface data integration. He earned his PhD degree at the Petroleum and Geosystems Engineering Department of UT Austin in 2013 and worked as a petrophysicist/rock physicist for BP America and BHP Billiton from 2013 to 2017. He cofounded and chaired the SPWLA PDDA SIG and initialized a student scholarship for PDDA-related graduate research. He has also served as an associate editor for several international scholastic journals, including *Petrophysics* and *SPE Reservoir Evaluation & Engineering*. He was selected to receive the Regional Formation Evaluation Technical Award by SPE – Gulf Coast (2018), the Meritorious Service Award (2019), the Outstanding Associate Editor Award (2020), the SPWLA Meritorious Technical Award (2021), and the Regional Data Science and Engineering Analytics Technical Award by SPE – Gulf Coast (2022).

**Michael Ashby** is a petrophysical advisor for Devon Energy. Prior to that, he was a staff petrophysicist with the Advanced Analytics and Emerging Technologies team at Anadarko Petroleum Corporation. Before joining Anadarko, Michael worked as a petrophysicist for Apache Corporation and Baker Hughes. He started his career as a wireline field engineer for SLB. Michael received his bachelor's degree in earth sciences from Edinboro University of Pennsylvania in 2005. He has been a member of SPWLA since 2008. Michael has served as the VP Downtown for the SPWLA Houston Chapter (2013–2014). He has also been on the Technology

Committee (2017). Michael has served on the board of the SPWLA PDDA SIG as the Chairman (2020–2021) and Vice Chairman (2019).



**Andrew McDonald** is currently the product manager for AI, Machine Learning, and Python at Geoactive Limited. He has over 17 years of industry experience, which spans petrophysics, geoscience, and machine learning. He currently provides petrophysical and machine-learning expertise to the development of interactive petrophysics. Andrew holds an MSc degree in earth science from the Open University and a BSc (Hons) degree in geology and petroleum geology from the University of Aberdeen. He has coauthored several technical papers and was selected as a Distinguished Speaker by the SPWLA in 2021.



**Wen Pan**, a data scientist at Shell since 2022, specializes in applying deep-learning techniques to enhance subsurface characterization and modeling. He also contributes as a guest editor and associate editor for the *Petrophysics* journal. Dr. Pan completed his petroleum engineering undergraduate degree at China University of Petroleum in 2016 and earned his PhD in petroleum engineering from The University of Texas at Austin in 2022. His research focuses on innovative machine-learning methods for subsurface applications, resulting in more than 10 journal papers and conference proceedings.



**Tianqi Deng** currently works as an interpretation development engineer at SLB. He earned his PhD in petroleum engineering from The University of Texas at Austin. His specialties include petrophysics and well-log interpretation, inverse problems, and machine learning.



**István Szabó** represents the Petrophysics Team as the discipline lead petrophysicist at MOL Group (Hungary). He has been working there for 17 years, except for a “short break” serving as a petrophysicist on the Geoscience Services Team of Maersk Oil UK (Aberdeen). By academic background, he holds an MSc degree in environmental engineering, geophysics and a BSc degree in engineering computer sciences. He is impassioned

by reverse engineering and data-driven solutions of well-log interpretation challenges.



**Pál Péter Hanzelik** is currently the lead data scientist of MOL Group. He graduated as a chemical engineer in 2014 and has worked on several artificial intelligence projects upstream, downstream, and wholesale. He led the MOL Group's data science team 1 year ago. His specialties include quality and quantitative predictive analytics, multivariate data analysis, and developing a time series forecasting model. He is enthusiastic about Industry 4.0 and artificial intelligence.



**Csilla Kalmár** is working as a data scientist at MA-Coding Ltd. She graduated as a geoinformatics specialist in 2012 from the University of Szeged and as a geophysicist engineer in 2014 from the University of Miskolc. She started her career at MOL Plc. as a geophysicist in the Group Exploration Team. In 2021, her scope of duties expanded, and she started performing data science tasks as well. She is one of the members of the data science team of MA-Coding Ltd. Her specialties include data preparation, multivariate data analysis, and time series forecasting. She is broadening her knowledge in the direction of data engineering tasks.



**Saleh Alatwah** is a reservoir petrophysicist and data scientist currently working in the Saudi Aramco Reservoir Characterization Department. He holds a master's degree in geo-energy engineering from TU Delft and has 10 years of experience in upstream operations, with a specialization in subsurface modeling. Saleh's expertise spans from exploration to production, encompassing both static and dynamic domains.



**Jaehyuk Lee** is an electrical engineering scientist working in Wireline Services for Oilfield Services and Equipment, Baker Hughes, since 2019. He has 15 years of research experience as an electrical engineer and a research scientist. He works on the development of innovative logging tools using various modalities such as an NMR, nuclear, and acoustics and also does research on digital signal processing and machine-learning methods for data analysis. He received a PhD degree in biotechnology from the University of Cambridge (UK) and MSc and BSc degrees, both in radio sciences and engineering from Korea University (South Korea).