PETROPHYSICS, VOL. 62, NO. 4 (AUGUST 2021); PAGES 393–406; 10 FIGURES; 2 TABLES. DOI:10.30632/PJV62N4-2021a4

Synthetic Sonic Log Generation With Machine Learning: A Contest Summary From Five Methods

Yanxiang Yu[1], Chicheng Xu[1], Siddharth Misra[1], Weichang Li[1], Michael Ashby[1], Wen Pan[2], Tianqi Deng[2], Honggeun Jo[2], Javier E. Santos[2], Lei Fu[3], Chengran Wang[3], Arkhat Kalbekov[4], Valeria Suarez[4], Epo Prasetya Kusumah[5], Mohammad Aviandito[5], Yogi Pamadya[5], and Hossein Izadi[6]

ABSTRACT

Compressional and shear sonic traveltime logs (DTC and DTS, respectively) are crucial for subsurface characterization and seismic-well tie. However, these two logs are often missing or incomplete in many oil and gas wells. Therefore, many petrophysical and geophysical workflows include sonic log synthetization or pseudo-log generation based on multivariate regression or rock physics relations. Started on March 1, 2020, and concluded on May 7, 2020, the SPWLA PDDA SIG hosted a contest aiming to predict the DTC and DTS logs from seven "easy-to-acquire" conventional logs using machine-learning methods (GitHub, 2020). In the contest, a total number of 20,525 data points with half-foot resolution from three wells was collected to train regression models using machine-learning techniques. Each data point had seven features, consisting of the conventional "easy-to-acquire" logs: caliper, neutron porosity, gamma ray (GR), deep resistivity, medium resistivity, photoelectric factor, and bulk density, respectively, as well as two sonic logs (DTC and DTS) as the target. The separate data set of 11,089 samples from a fourth well was then used as the blind test data set. The prediction performance of the model was evaluated using root mean square error (RMSE) as the metric, shown in the equation below:

$$RMSE = sqrt(\frac{1}{2} * \frac{1}{m} * \left[\sum_{i=1}^{m} (DTC_{pred}^{i} - DTC_{true}^{i})^2 + (DTS_{pred}^{i} - DTS_{true}^{i})^2\right]$$

In the benchmark model, (Yu et al., 2020), we used a Random Forest regressor and conducted minimal preprocessing to the training data set; an RMSE score of 17.93 was achieved on the test data set. The top five models from the contest, on average, beat the performance of our benchmark model by 27% in the RMSE score. In the paper, we will review these five solutions, including preprocess techniques and different machine-learning models, including neural network, long short-term memory (LSTM), and ensemble trees. We found that data cleaning and clustering were critical for improving the performance in all models.

INTRODUCTION

Well logging is essential for the oil and gas industry to understand the in-situ subsurface petrophysical and geomechanical properties (Xu et al., 2019). Certain well logs, like GR, resistivity, density, and neutron, are considered as "easy-to-acquire" conventional well logs and are deployed in most wells. Other well logs, like nuclear magnetic resonance (NMR), dielectric dispersion, elemental spectroscopy, and dipole/shear sonic, are deployed in a limited number of wells. Easy-to-acquire well logs can be processed using statistical and machine-learning methods to synthesize the well logs that are not frequently acquired in each well. Researchers have explored the possibility of synthesizing certain "hard-to-acquire" well logs under data constraints (Tariq et al., 2016; Li and Misra, 2017; He and Misra, 2019).

Sonic logging tools transmit compressional and shear waves through a geological formation. These waves are sensitive to the mechanical properties of the formation, such as Young's modulus and Poisson's ratio. Compressional and shear traveltime logs (DTC and DTS, respectively) can be

Manuscript received by the Editor March 1, 2021; revised manuscript received May 11, 2021; manuscript accepted May 19, 2021.
[1]SPWLA PDDA SIG; Shell International Exploration and Production Inc., yueureka@gmail.com; Aramco Americas: Aramco Research Center - Houston, Chicheng.Xu@aramcoamericas.com; Texas A&M Unversity, misra@tamu.edu; Aramco Americas: Aramco Research Center - Houston, Weichang.Li@aramcoamericas.com; Devon Energy, ashby149@aol.com
[2]Team UTFE; University of Texas at Austin, wenpan@utexas.edu; tianqizx@utexas.edu; honggeun.jo@utexas.edu; jesantos@utexas.edu
[3]Team iwave; Aramco Americas: Aramco Research Center - Houston, lei.fu.rice@gmail.com; University of Houston, wangchengran123@gmail.com
[4]Team RockAbuser; Colorado School of Mines, akalbekov@mines.edu; vasuarezbolivar@mines.edu
[5]Team SedStrat; Universitas Pertamina, epo.pk@universitaspertamina.ac.id; aviandito@gmail.com; yogipamadya@gmail.com
[6]Team Ipetro; University of Alberta, hosizadi@ualberta.ca

computed from the waveforms recorded by the receivers of the sonic logging tool. Compressional waves travel through both the rock matrix and pore-filling fluid, while shear waves travel only through the rock matrix. The wave traveltime depends on the porosity, consolidation, and elastic moduli, as well as the composition and microstructure of the rock. The DTC and DTS contain information about the formation porosity, rock brittleness, and Young's modulus, to name a few. However, a sonic logging tool may not always be available due to financial or operational constraints. Many empirical models have been developed in the oil and gas industry to predict the sonic traveltimes, especially DTS, from other well logs. Researchers have predicted DTS or DTC logs using empirical physics-based equations (Iverson and Walker, 1988; Greenberg and Castagna, 1992), empirical correlations (Maleki et al., 2014), and self-similar models (Keys and Xu, 2002). Other researchers predicted DTS and DTC logs in thin beds using petrophysical properties instead of raw conventional logs, (e.g., Baines et al., 2008). Our paper summarizes a few robust workflows to synthesize both DTC and DTS logs from easy-to-acquire conventional well logs using simple machine-learning models.

Several studies have implemented machine-learning techniques to determine the sonic logs from other well logs. Artificial neural network (ANN), adaptive data-driven inference system, and support vector machines were used to predict both compressional and shear sonic traveltimes from GR, bulk density, and neutron porosity (Elkatatny et al., 2016). The study achieved a correlation coefficient of 0.99 when tested on field data. In another study, shear wave velocity (reciprocal of DTS) is predicted using the intelligent system, which combined the algorithms of fuzzy logic, neuro-fuzzy, and ANNs. The mean squared error during the testing stage was around 0.05 (Rezaee et al., 2007). A similar study applied a committee machine with intelligent systems to predict sonic traveltime from conventional well logs (Asoodeh and Bagheripour, 2012).

Tariq et al. (2016) developed an artificial neural network model to determine sonic traveltime by processing gamma ray, bulk density, and neutron porosity. Sonic traveltime was predicted with an R-squared of 0.96. Comparative study of various machine-learning methods for purposes of sonic-log synthesis was extensively tested by He et al. (2018). Following that, He et al. (2019) proposed a reliable workflow that can synthesize the DTC and DTS logs, as well as generate a reliability indicator for the synthesis logs to help the user better understand the performance of the shallow-learning models during deployment in new wells. Bader et al. (2018) proposed a method to use density and neutron logs for estimating the missing sonic logs using Bayesian estimation. There have been new data-driven models for the purpose of sonic-log prediction (Onalo et al., 2018; Onalo et al., 2019). In a more recent study, Pham et al. (2020) used a bidirectional convolutional long short-term memory (bidirectional ConvLSTM) cascaded with fully connected neural networks to predict missing sonic well logs from gamma ray, density, and neutron-porosity logs.

Readers should note that the excellent predictive performances of the data-driven methods described in the previous paragraphs are specific to the data set used, the location/event of interest, and the task to be performed. These data-driven methods need to be retrained and re-evaluated when there is a change in data, location, event, or task. The data-driven methods need data that have high quality and large size for achieving the high-predictive performance. The performance of data-driven methods described in the previous paragraph cannot be achieved by classical physics-based methods. However, classical methods tend to be more generalizable as compared to the data-driven methods, which are suited for specific data, task, event, and location. Validation of classical physics-based techniques needs smaller-sized data sets as compared to the abovementioned data-driven methods.

There has been an increasing excitement about applying machine-learning and artificial intelligence (AI) methods in the oil and gas industry. On those lines, the sonic log synthesis or prediction by processing conventional logs using machine-learning techniques is a perfect demonstration of the power of machine-learning applications. Many free and open-source packages now exist that provide powerful additions to the petrophysists' or rock physicists' toolbox. One of the best examples is scikit-learn (http://scikit-learn.org/), a collection of tools for machine learning in Python to complete the machine-learning process for this problem. In this tutorial, we'll be using functions from this library and provide a machine-learning workflow to predict the DTC and DTS logs by processing conventional logs. The prediction models are trained by processing data from the concatenation of logs from three similar wells and use feature sets derived from the seven conventional logs—caliper, neutron, GR, deep resistivity, medium resistivity, photoelectric factor, and density—and then the model is used to generate the two targets' DTC and DTS logs in another similar well. The predicted values are saved in the same format as the given sample_submission.csv and submitted together with a notebook for judgment. In the rest of the paper, we will review the five solutions from the contest, which were developed by Team UTFE, Team iwave, Team RockAbuser, Team SedStrat, and Team Ipetro. The data preprocessing techniques and machine-learning models from the solutions will be discussed in detail. Then, we will present a comparison of their prediction results and evaluate the synthetic log generation performance from the petrophysical point of view.

DATA PREPARATION AND PREPROCESSING

Data Preparation

In this paper, four wells from the Volve data set were selected to train and test the machine-learning algorithm. The Volve data set was released by the Equinor Petroleum Company, which disclosed all subsurface and operating data from the Volve Field—located 200 kilometers west of Stavanger at the southern end of the Norwegian sector.

We selected four wells from the data set: Well A, Well BT2, Well F11-A, and Well F-1A. Each of the four wells has data from nine well logs aligned by depth—where caliper, neutron, GR, deep resistivity, medium resistivity, photoelectric factor, and density logs are used as the input features, and DTC and DTS logs are the output. The four wells are close to each other with similar geology, and the data are relatively complete. We then purposely removed the measured depth information of each well and concatenated the three training wells to one data set in order to avoid any data leaking issues. Well A, Well BT2, and Well F11-A are used for training the model, and Well F-1A is used for testing.

Data Preprocessing and Feature Engineering

Good data quality control and extraction of useful information from the features are key to the success of machine-learning models. Some basic steps include removing the bad hole intervals and absent NaN-values in the logs, taking the logarithm scale of resistivity logs, and normalizing the input logs by either standard normalization to Gaussian distribution of 0 mean and 1.0 standard deviation, or min-max normalization to 0 and 1.0. Beyond these, some more advanced techniques incorporate petrophysics, and signal processing methods are also very useful to extract useful information from the logs.

For example, Team UTFE applied three key techniques: 1. Median filters were used for the input logs to alleviate aliasing problems caused by data interpolation and eliminate outliers; 2. Gradients of different logs at adjacent depths were generated to take the local heterogeneity into consideration; and 3. Separate the training data into different zones based on the GR response and the physical depth of the training data set and trained five different ANN models that were used to predict the test data separately. Fig. 1 shows the zonation performed for different wells based on the GR logs.

Team iwave focused on removing anomalies, choosing a good combination of features, and creating an extra feature by labeling the lithology. They used a support vector machine (SVM) classifier to detect the anomalous data points. In addition, based on their analysis of the relation between DTC and DTS, the ratio DTC/DTS was added as a new feature to represent different lithology.

Team RockAbusers conducted extra analysis on the output logs. They found that the top and bottom zones of the DTS log change dramatically and therefore cause the largest error. Thus, they first used different feature sets for the top and bottom zones as these sections contain some hydrocarbons based on crossovers in the density-neutron curves and secondly added resistivity to the features list as it may be contributing more over these zones.

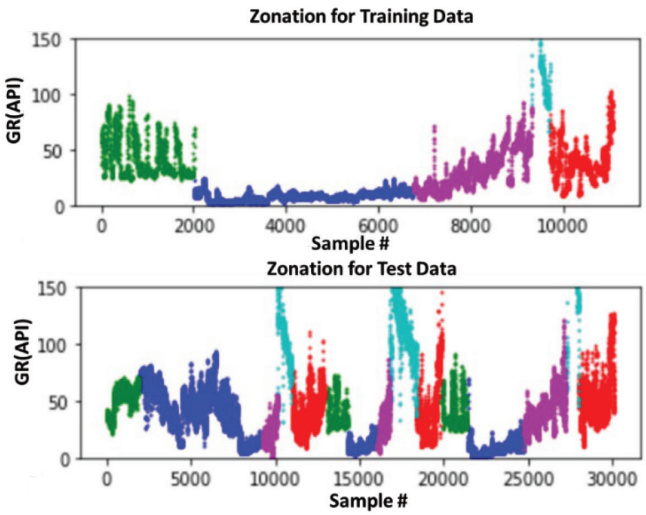

Fig. 1—Zonation performed for different wells based on GR logs.

Team SedStrat created new features based on the petrophysics domain knowledge. They realized that DTC and DTS would behave differently in different lithology, fluid, and borehole conditions. Therefore, discriminants were created to differ lithology, fluid, and borehole conditions. They eventually found that lithology was one of the most useful features for predictions. The lithology feature is based on the classical neutron-density (NPHI-RHOB) crossplot used in conventional petrophysical analysis. To create the lithology label, they started with points of NPHI-RHOB observations and sandstone, limestone, and dolomite matrix lines and then formulated a simple geometry problem of distance between the observation points and the matrix line. The nearest matrix line from that point is the lithology of the formation. Lastly, the resulting lithology with the standard DTC value of the matrix was encoded. Fig. 2 shows the lithology creation results.

Team Ipetro conducted clustering methods to separate the training data set into different clusters for training. Two different clustering methods were applied: incremental clustering (Izadi et al., 2020) and clustering based on

prediction results of validation databases, respectively. The clustering results using the incremental clustering algorithm are plotted in Fig. 3; however, the V_s prediction did not meet the expectation under this clustering method. The crossplot for V_s prediction in the validation database is presented in Fig. 4. There is an underestimation for intervals corresponding to high GR and neutron values. The database can be divided into two clusters: less than 230 and more than 230.

In summary, the domain knowledge in petrophysics is very helpful in processing the logs and extracting features. The advanced techniques, such as creating zonation, creating new features and labels, and clustering the data set into different groups for training, greatly helped to improve the model performance.

MACHINE-LEARNING MODELS

A number of different machine-learning models were chosen by the participating teams, including ensemble models, tree-based models, ANN, as well as deep-learning recurrent neural network (RNN) models, such as the LSTM network. Combined with various data cleaning and preprocessing, and sometimes with domain prior information or models, these methods all demonstrated relatively good performance in this particular application.

Team UTFE adopted an interesting zone-specific training and prediction strategy, which works quite successfully in this case. After preprocessing to remove aliasing problems and eliminate outliers, zonation was performed based on GR logs for both wells. As a result, depth samples from each well log (including the testing well) were partitioned into five different depth zones. After that, a basic ANN model was adopted for all zones. However, training was conducted for each zone separately. The five trained models were then applied to predict sonic logs at respective zones of the testing well. With samples within each zone being presumably more homogeneous, the ANN model structure was of relatively low complexity, containing two hidden layers. The first layer has 24 neurons, the second has 12, and the output layers are DTS and DTC, respectively. The activation function for the first two layers is the Rectified Linear Unit (ReLU), and the output layer's activation function is sigmoid. The input layer takes a vector of three consecutive depth samples of input log data as input to account for spatial correlation. Min-max scaling was used to scale both the features and target data due to the sigmoid function used in the output layer. The low-prediction error and high-predictive performance achieved by the team indicates a powerful demonstration of the value

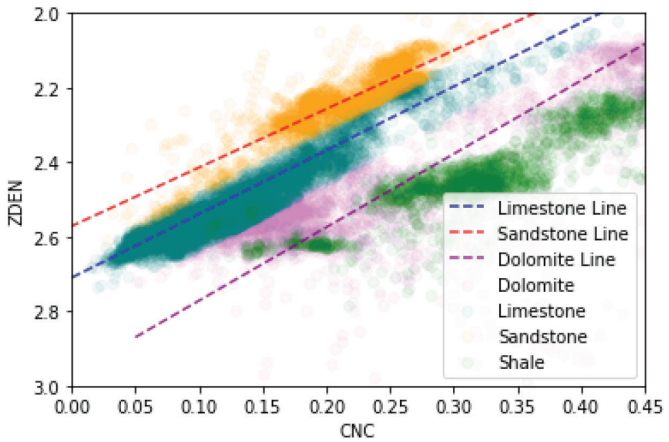

Fig. 2—Lithology label creation based on density and caliper log.

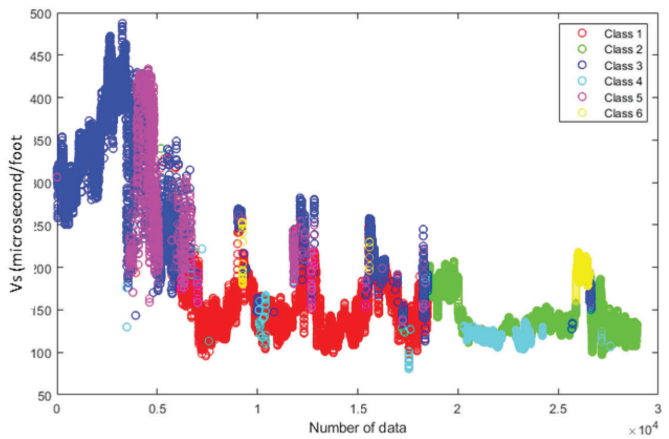

Fig. 3—Clustering results using an incremental clustering algorithm. Based on these classes, V_s prediction failed.

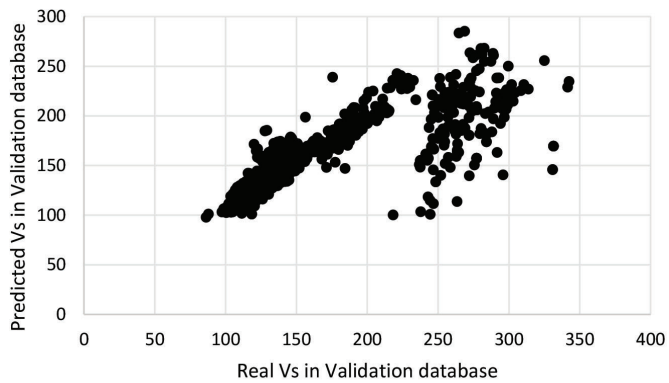

Fig. 4—Crossplot for V_s prediction in the validation database. There is an underestimation for intervals corresponding to high GR and neutron values. The database can be divided into two clusters: less than 230 and more than 230.

of domain prior, which enables an efficient low-complexity network model. However, this strategy might arguably require repeated zonation and retraining for predicting on new wells and may run into performance issues when either the zonation of the testing well is not necessarily consistent with the training well or, in a mild case, has very imbalanced zonal distribution.

Team iwave used an LSTM network model (Gers et al., 1999). The basic architecture of bidirectional LSTM consists of two hidden layers of opposite directions to the same output. This type of structure is capable of learning bidirectional order dependence in sequence-prediction problems. An LSTM cell can learn to recognize an important input with an input gate, store it in the long-term state, learn to preserve it for as long as it is needed, ensure that it is maintained in the forget gate, and learn to extract it whenever it is needed. This greatly helps the model capture long-term sequence patterns and has been quite successful in learning time series, text sequences, and audio clips. It is well suited to tasks like prediction with sequence data. Because the properties measured by various well-log techniques are influenced by properties from the adjacent depth of both upside and downside, this form of sequence deep-learning model can be a good fit for the well-log prediction problem. Before feeding the log data into the LSTM model, significant effort was made to remove anomalies and choose a good combination of features, as well as experiment with different hyperparameters during model training, including input sequence length, number of hidden layers, number of nodes, batch size, dropout ration, regularization weighting, learning rate, and early stop (number of epochs). The LSTM model was trained on a single training set, and the training model was applied to the entire test well with very good predictive performance. This is a case that successfully demonstrated the power of a relatively deep network model with a structure matching the nature of the data and the problem setting. As it is entirely data driven without requiring zonation, this might have more generalization advantage in practice.

A related but different model, Elman neural network (Elman, 1990), was adopted by Team Ipetro. Elman is one type of RNN model. It tries to model the sequence dependency by feeding back the hidden state to be combined with the new input. However, it is generally known that the Elman model can suffer from the so-called vanishing/exploding gradients problem associated with the activation function and the eigenvalues of the weight matrix. This can hurt its ability to learn long-range dependencies. In comparison, the forget gate in LSTMs allows it to delay this problem and hence process longer input sequences.

A Random Forest (RF) regression model was implemented by Team RockAbusers, which also achieved very good sonic log prediction. RF is an ensemble-learning technique (Breiman, 2001). It uses multiple classification and regression trees (CARTS) and a technique called Bootstrap Aggregation to train each decision tree on different data samples where sampling is done with replacement and aggregates the mean prediction of individual trees. This ensures that the ensemble model does not rely too heavily on any individual feature and makes fair use of all potentially predictive features. It also further prevents overfitting by letting each tree draw a random sample from the training data when generating its splits. Another benefit of using Random Forest is that it outputs a hierarchy of important features. For instance, to predict the sonic porosity tool results, the other porosity tools (density neutron) had the highest importance.

Team SedStrat also adopted the ensemble method in their submission for predicting DTC. Specifically, a two-step process was applied. The first step is prediction using an ensemble of popular machine-learning algorithms, namely Random Forest, GBR, AdaBoost, ExtraTree, Lasso Regression, Ridge Regression, and Linear Regression. Results from these algorithms will then be used as features for the second step prediction. The model that is being used to wrap the ensemble is LightGBM (Ke et al., 2017) in the second step. Prediction for DTS was directly implemented using LightGBM, as the ensemble method was found to perform not as well, according to the team.

In summary, the model developed by each team shows their pro and cons. For example, the UTFE model takes the benefit of prior-zonation-based approach in simplifying the model complexity, with the cost of extra effort in exploring the group-specific homogeneity and reducing the potential zonal imbalance. In contrast, the Team iwave model, the simpler workflow with a single model, comes at the cost of training a more complex network architecture and the potential performance difference when applied to various formation and rock types. Similar remarks have also been provided for the models adopted by the other teams.

RESULTS COMPARISON

We plotted the predicted results in the test well submitted by all five teams. Both log tracks and crossplots are shown for better view and comparison. Different colors are used to differentiate the teams, as shown in Table 1.

Table 1—Colors of DTC and DTS Logs for Each Team

Team	Color of DTC Log	Color of DTS Log
UTFE	Blue	Blue
iwave	Black	Black
RockAbuser	Red	Red
SedStrat	Green	Green
Ipetro	Orange	Orange
Original Logs	Fuchsia	Purple

Table 2—The RMSE and R2 Achieved by Each Team in the Contest

		UTFE	iwave	RockAbuser	SedStrat	Ipetro
RMSE	**DTC**	4.8598	5.7163	4.6475	4.4558	4.7402
	DTS	16.7897	16.8055	18.0997	19.0673	21.2382
R2	**DTC**	0.8875	0.8443	0.8971	0.9054	0.8929
	DTS	0.8569	0.8566	0.8337	0.8154	0.7710

Table 2 shows the RMSE and R2 achieved by each team in the contest. All teams have outperformed the benchmark set by the committee. An interesting thing to note is that the team that achieved the best RMSE in DTC prediction may not necessarily achieve the best in DTS prediction.

Figures 5 and 6 show the comparison of results in a log profile over the shallow interval and deep interval, respectively. As expected, the overall prediction performance of DTC is better than DTS. Both DTC and DTS predictions are much better in the deep interval, possibly due to less variability or more homogeneity. In the shallow interval, there are multiple zones showing significantly underestimation or overestimation when the sonic logs exhibit peak or trough values. We also notice that the shallow section's predictions are systematically faster than the measured values, while the deep section's predictions are systematically slower. This is a common issue with statistical regression methods. It indicates the model complexity is still low in capturing maximum variabilities.

Figures 7 to 10 show the crossplots over certain intervals for both DTC and DTS, where the top left figures are the combination of all five teams, and the color of each plot is corresponding to Table 1. The top middle plots are the results of Team UTFE, the top right plots are the results of Team SedStrat, the bottom left plots are the results of Team RockAbuser, the bottom middle plots are the results of Team iwave, and the bottom right plots are the results of Team Ipetro.

Figures 7 and 8 show the comparison of results in a crossplot over the shallow interval for DTC and DTS, respectively. We can see while the points of DTC logs are reasonably distributed along the $y = x$ line, many points of DTS logs are far off the $y = x$ trend (more points below the $y = x$ line). Overall, the prediction of both DTC and DTS are poor, and we doubt that the predicted results can be used in any petrophysical or geophysical workflows.

Figures 9 and 10 show the comparison of results in a crossplot over the deep interval for DTC and DTS, respectively. In this interval, points of both DTC and DTS logs are close to the $y = x$ line. However, the predicted DTS logs from all teams are slightly higher than the original logs (above the $y = x$ line). This is possibly due to the averaging effect of statistical regression methods. Overall, the prediction of both DTC and DTS are considered good enough to be used in petrophysical or geophysical workflows.

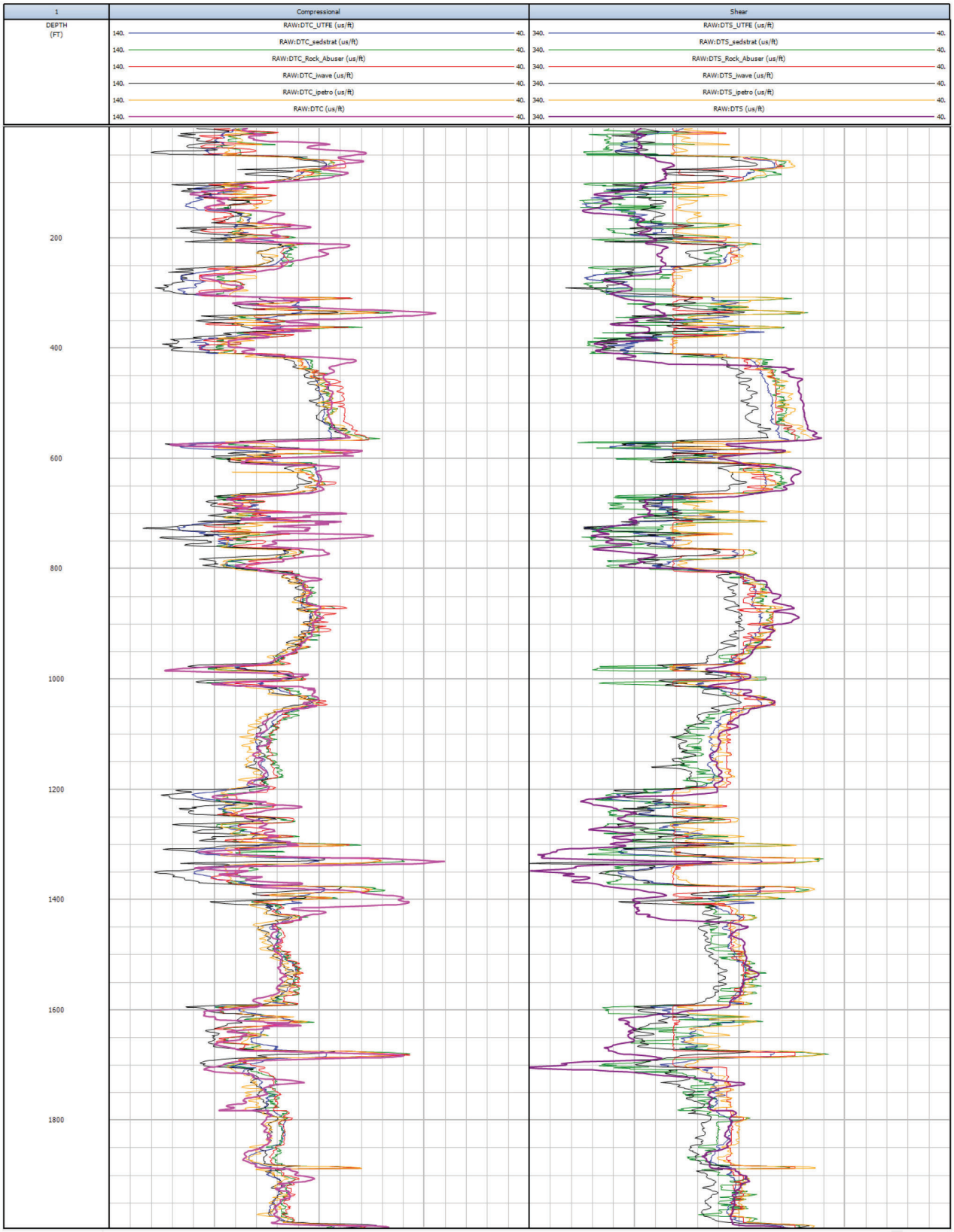

Fig. 5—Comparison of all winning teams in a log profile over the shallow interval: (a) DTC and (b) DTS.

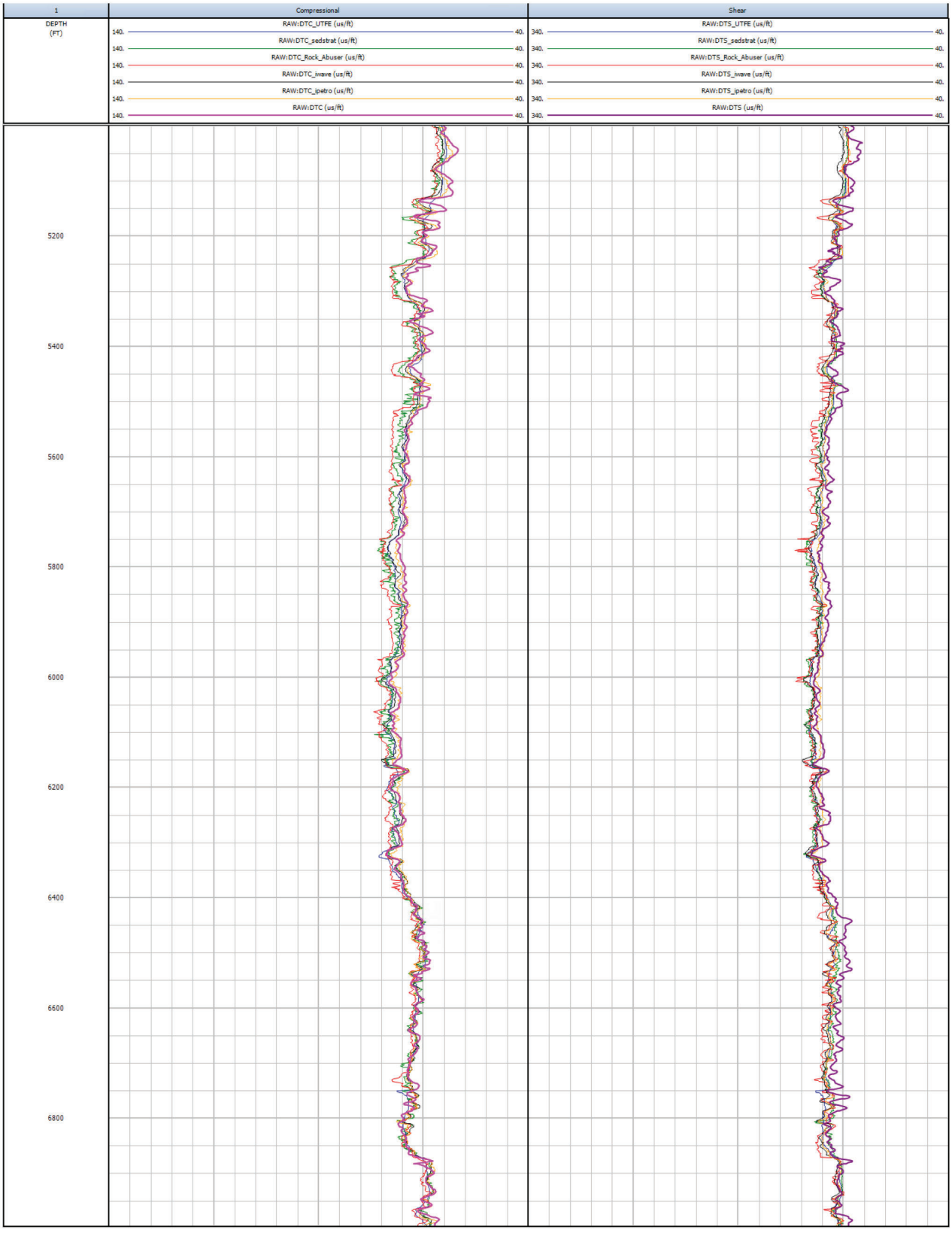

Fig. 6—Comparison of all winning teams in a log profile over the deeper interval: (a) DTC and (b) DTS.

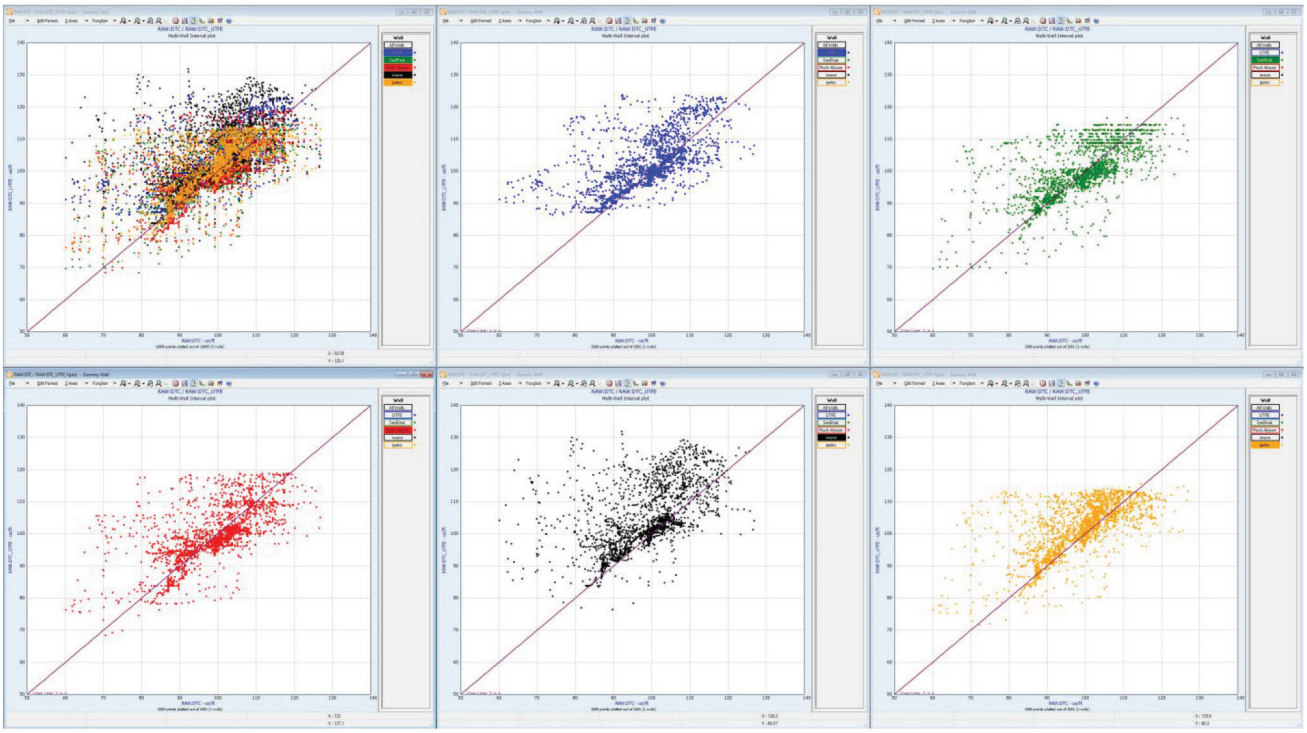

Fig. 7—DTC comparison of all winning teams in the crossplot space over the shallow interval. Original DTC: *x*-axis, Predicted DTC: *y*-axis.

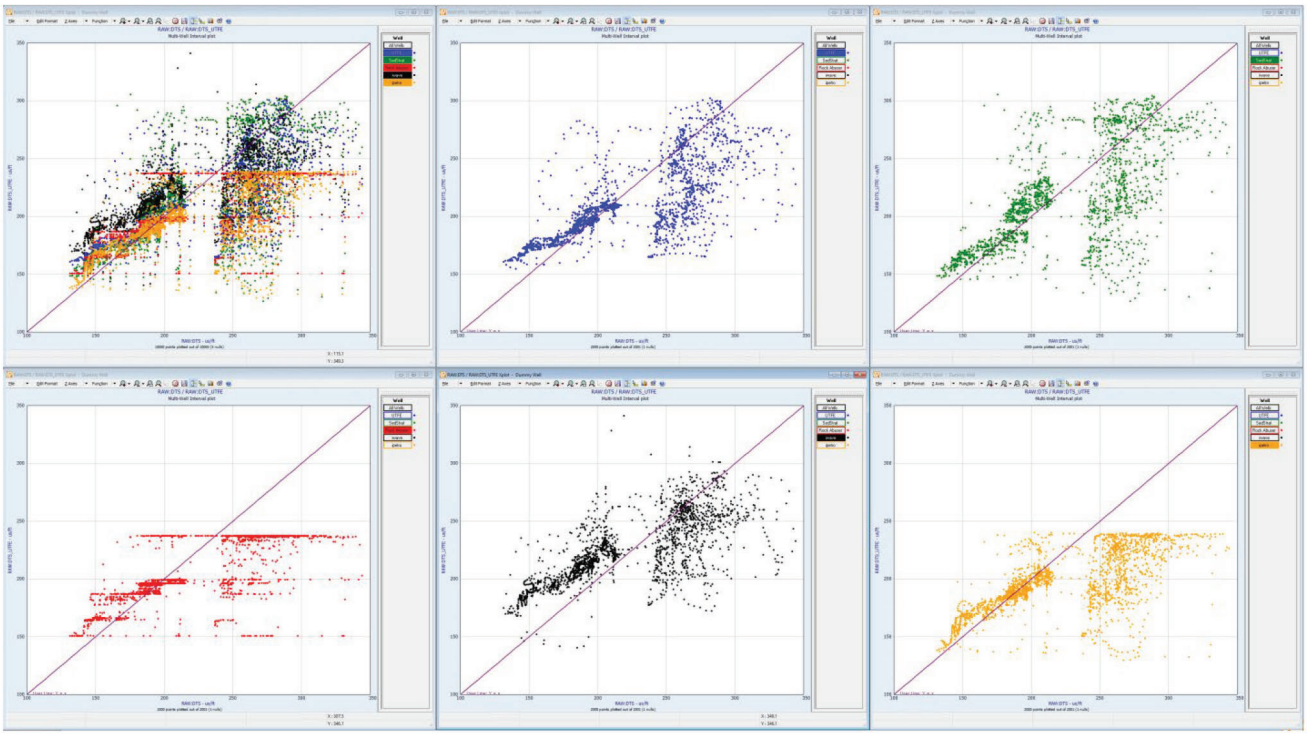

Fig. 8—DTS comparison of all winning teams in the crossplot space over the shallow interval. Original DTS: *x*-axis, Predicted DTS: *y*-axis.

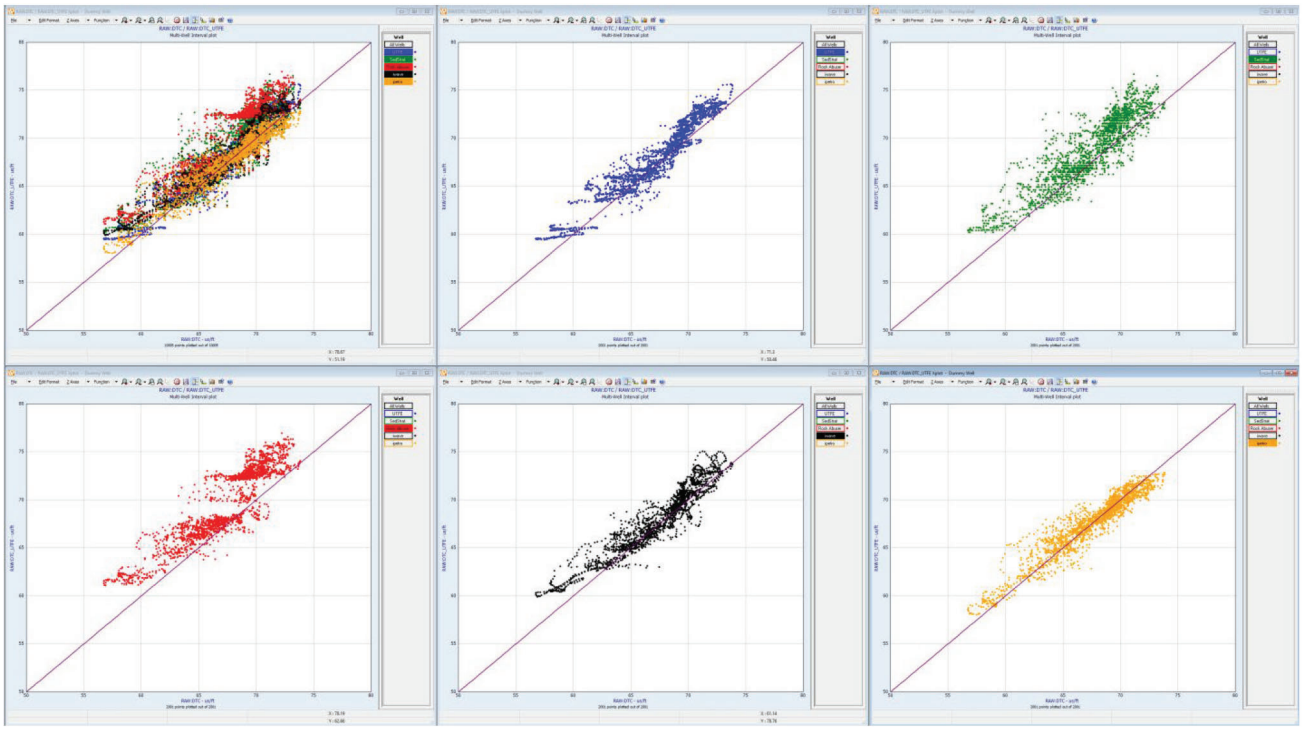

Fig. 9—DTC comparison of all winning teams in the crossplot space over a deeper interval. Original DTC: *x*-axis, Predicted DTC: *y*-axis.

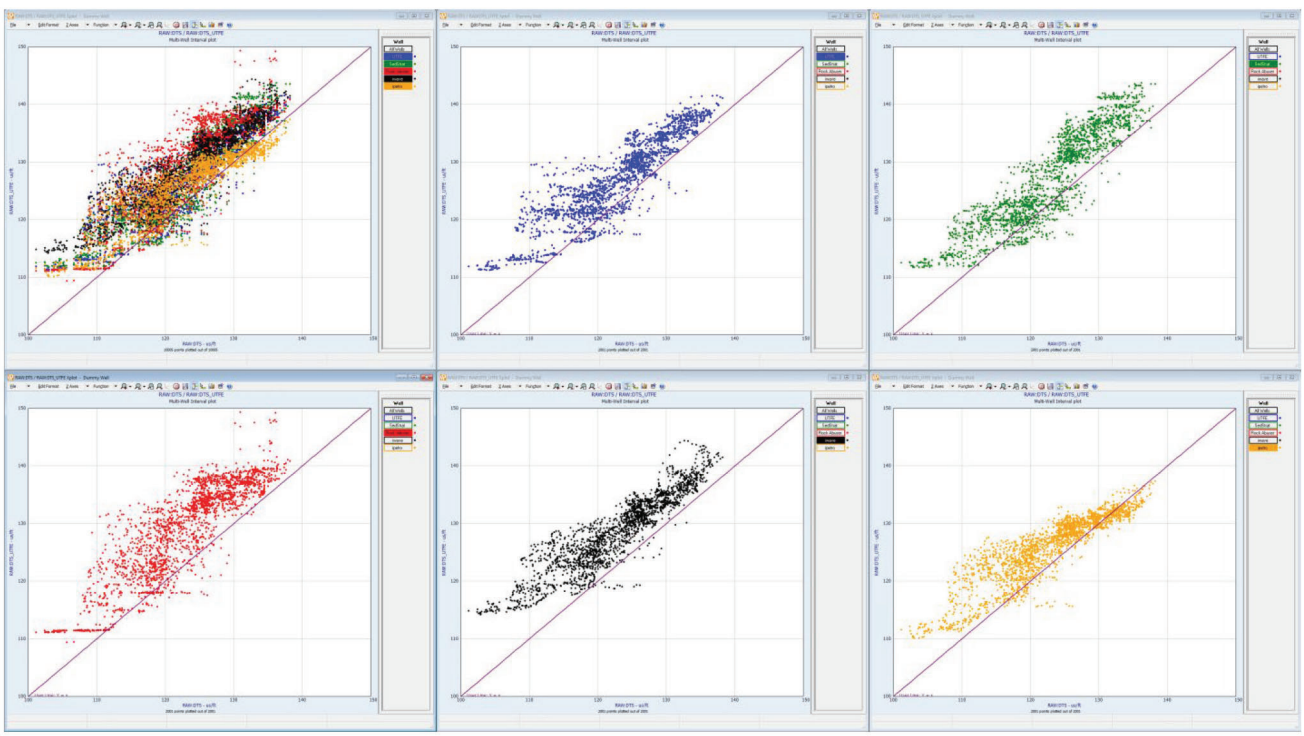

Fig. 10—DTS comparison of all winning teams in the crossplot space over a deeper interval. Original DTS: *x*-axis, Predicted DTS: *y*-axis.

SUMMARY

This contest demonstrated a distributed and collaborative technology development effort. Over a two-month span, teams from all over the world worked diligently on improving their models. We saw some great results and a significant improvement in their models' performance compared to the benchmark model.

The teams were able to demonstrate their machine-learning workflow on a practical petrophysical problem: data set preparation and quality assurance, feature engineering with outlier handling and clustering, training and testing a regression model, and finally, blind-testing (similar to the real-world deployment) the model on the hidden data set. Various models have been adopted by the different teams. We found that for this particular petrophysical problem, where the data set was relatively small (with a training data set of ~20,000 samples from only three wells), the model itself might not be the key to the success of predicting on the new data set, but rather many other methods that are applied to improve the performance and stability of the model, such as make special treatments for the anomalies and outliers, train different models for zones that show a very distinct DTC/DTS range, train multiple regression models, and/or combine them.

We observed that some modeling inconsistency could also be due to borehole quality and poor raw measurements in both the training and testing data. These can result in erroneous well-log responses that do not represent the formations. Being able to predict these logging errors does not aid formation evaluation. Caution must be exercised to minimize these artifacts before starting training and testing. As petrophysicists, we want to be able to use the data to describe the rocks, pores, and fluids accurately.

The proprietary nature of the oil and gas industry, in general, limited many machine-learning methods to be adopted in the petrophysical domain. With open data sets becoming more readily available, we hope this contest provides an example of the enthusiasm and talent to help build up a shared knowledge base of the industry.

ACKNOWLEDGMENTS

A note of thanks goes to SparkCognition for sponsoring the event and to Equinor for releasing the Volve data set. We also thank the other members of the SPWLA PDDA SIG ML Contest Committee: Brendon Hall, Bin Dai, Zheng Gan, and Yan Xu for their contributions.

NOMENCLATURE

Abbreviations

AI =	artificial intelligence
ANN =	artificial neural network
CARTS =	classification and regression trees
DTC =	compressional sonic traveltime logs
DTS =	share sonic traveltime logs
GR =	gamma ray
LSTM =	long short-term memory
NMR =	nuclear magnetic resonance
ReLU =	rectified linear unit
RF =	Random Forest
RMSE =	root mean square error
RNN =	recurrent neural network
SVM =	support vector machine

REFERENCES

Asoodeh, M., and Bagheripour, P., 2012, Prediction of Compressional, Shear, and Stoneley Wave Velocities From Conventional Well Log Data Using a Committee Machine With Intelligent Systems, *Rock Mechanics and Rock Engineering*, **45**(1), 45–63. DOI: 10.1007/s00603-011-0181-2.

Bader, S., Spikes, K., and Fomel, S., 2018, Missing Well-Log Data Prediction Using Bayesian Approach in the Relative-Geologic Time Domain, *SEG Technical Program Expanded Abstracts 2018*, 804–808. DOI: 10.1190/segam2018-2997278.1.

Baines, V., Bootle, R., Pritchard, T., Macintyre, H., and Lovell, M.A., 2008, Predicting Shear and Compressional Velocities in Thin Beds, Paper I, *Transactions*, *SPWLA 49th Annual Logging Symposium*, Austin, Texas, USA, 25–28 May.

Breiman, L., 2001, Random Forests, *Machine Learning*, **45**, 5–32. DOI: 10.1023/A:1010933404324.

Elkatatny, S.M., Zeeshan, T., Mahmoud, M., Abdulazeez, A., and Mohamed, I.M., 2016, Application of Artificial Intelligent Techniques to Determine Sonic Time From Well Logs, Paper ARMA-2016-755 presented at the 50th US Rock Mechanics/Geomechanics Symposium, Houston, Texas, USA, 25–29 June.

Elman, J.L., 1990, Finding Structure in Time, *Cognitive Science*, **14**(2), 179–211. DOI: 10.1016/0364-0213(90)90002-E.

Gers, F.A., Schmidhuber, J., and Cummins, C., 1999, Learning to Forget: Continual Prediction With LSTM, *Proceedings*, *Ninth International Conference on Artificial Neural Networks ICANN 99*, **470**, 850–855. DOI: 10.1049/cp:19991218.

Github website for SPWLA ML Contest, 2020, URL: https://github.com/pddasig/Machine-Learning-Competition-2020. Accessed June 1, 2021.

Greenberg, M., and Castagna, J., 1992, Shear-Wave Velocity Estimation in Porous Rocks: Theoretical Formulation,

Preliminary Verification and Applications, *Geophysical Prospecting*, **40**(2), 195–209. DOI: 10.1111/j.1365-2478.1992.tb00371.x.

He, J., Misra, S., and Li, H., 2018, Comparative Study of Shallow Learning Models for Generating Compressional and Shear Traveltime Logs, *Petrophysics*, **59**(06), 826–840. DOI: 10.30632/PJV59N6-2018a7.

He, J., Li, H., and Misra, S., 2019, Data-Driven In-Situ Sonic-Log Synthesis in Shale Reservoirs for Geomechanical Characterization, Paper SPE-191400, *SPE Reservoir Evaluation & Engineering*, **22**(4), 1225–1239. DOI: 10.2118/191400-PA.

He, J., and Misra, S., 2019, Generation of Synthetic Dielectric Dispersion Logs in Organic-Rich Shale Formations Using Neural-Network Models, *Geophysics*, **84**(3), D117–D129. DOI: 10.1190/geo2017-0685.1.

Iverson, W.P., and Walker, J.N., 1988, Shear and Compressional Logs Derived From Nuclear Logs, *SEG Technical Program Expanded Abstracts 1988*, 111–113.

Izadi, H., Sadri, J., Hormozzade, F., and Fattahpour, V., 2020, Altered Mineral Segmentation in Thin Sections Using an Incremental-Dynamic Clustering Algorithm, *Engineering Applications of Artificial Intelligence*, **90**, 103466. DOI: 10.1016/j.engappai.2019.103466.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., 2017, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems (NIPS 2017)*, **30**, 3146–3154. URL: https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. Accessed June 6, 2021.

Keys, R.G., and Xu, S., 2002, An Approximation for the Xu-White Velocity Model, *Geophysics*, **67**(5), 1406–1414. DOI: 10.1190/1.1512786.

Li, H., and Misra, S., 2017, Prediction of Subsurface NMR T2 Distribution From Formation-Mineral Composition Using Variational Autoencoder, *SEG Technical Program Expanded Abstracts 2017*, 3350–3354. DOI: 10.1190/segam2017-17798488.1.

Maleki, S., Moradzadeh, A., Riabi, R.G., Gholami, R., and Sadeghzadeh, F., 2014, Prediction of Shear Wave Velocity Using Empirical Correlations and Artificial Intelligence Methods, *NRIAG Journal of Astronomy and Geophysics*, **3**(1), 70–81. DOI: 10.1016/j.nrjag.2014.05.001.

Onalo, D., Adedigba, S., Khan, F., James, L.A., and Butt, S., 2018, Data Driven Model for Sonic Well Log Prediction, *Journal of Petroleum Science and Engineering*, **170**, 1022–1037. DOI: 10.1016/j.petrol.2018.06.072.

Onalo, D., Oloruntobi, O., Adedigba, S., Khan, F., James, L., and Butt, S., 2019, Dynamic Data Driven Sonic Well Log Model for Formation Evaluation, *Journal of Petroleum Science and Engineering*, **175**, 1049–1062. DOI: 10.1016/j.petrol.2019.01.042.

Pham, N., Wu, X., and Zabihi Naeini, E., 2020, Missing Well Log Prediction Using Convolutional Long Short-Term Memory Network, *Geophysics*, **85**(4), WA159–WA171. DOI: 10.1190/geo2019-0282.1.

Rezaee, M.R., Ilkhchi, A.K., and Barabadi, A., 2007, Prediction of Shear Wave Velocity From Petrophysical Data Utilizing Intelligent Systems: An Example From a Sandstone Reservoir of Carnarvon Basin, Australia, *Journal of Petroleum Science and Engineering*, **55**(3–4), 201–212. DOI: 10.1016/j.petrol.2006.08.008.

Tariq, Z., Elkatatny, S., Mahmoud, M., and Abdulraheem, A., 2016, A New Artificial Intelligence Based Empirical Correlation to Predict Sonic Travel Time, Paper IPTC-19005 presented at the International Petroleum Technology Conference, Bangkok, Thailand, 14–16 November. DOI: 10.2523/IPTC-19005-MS.

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.

Yu, Y., Misra, S., Oghenekaro, O., and Xu, C., 2020, Pseudosonic Log Generation With Machine Learning, A Tutorial for the 2020 SPWLA PDDA SIG ML Contest, *SPWLA Today*, **3**(2), 97–101. URL: https://www.spwla.org/Documents/Newsletter/SPWLA_Today_Newsletter_Vol3_No2.pdf. Accessed June 6, 2021.

ABOUT THE AUTHORS

Yanxiang Yu is currently working at Shell as an artificial intelligent resident focusing on machine-learning modeling for upstream applications. Previously, he worked as a senior research scientist at GOwell International for 5 years, where he focused on developing well-integrity evaluation logging tools. Yanxiang obtained his master's degree from the University of Houston in 2014. Since 2019, he has been serving as the Chair for SPWLA PDDA SIG and previously served as the Secretary of the SPWLA Resistivity SIG from 2017 to 2019.

Chicheng Xu is working at the Aramco Houston Research Center as a research petrophysicist focusing on petrophysics data-driven analytics that uses advanced computational techniques and artificial intelligence/machine learning for interpretation, classification, and modeling based on multiscale data integration. He obtained his PhD degree in petroleum engineering from the University of Texas at Austin in 2013 and had previously worked in various technology departments of Schlumberger, BP, and BHP Billiton. Chicheng has published more than 30 technical papers and received many technical awards from SPWLA and SPE.

Siddharth Misra is an associate professor in the Harold Vance Department of Petroleum Engineering, Texas A&M University, with a courtesy appointment in the Department of Geology and Geophysics, Texas A&M University. He is a researcher and educator in the field of formation evaluation, petrophysics, data analytics, and machine learning. Siddharth has advanced new machine-learning procedures for the interpretation of geophysical subsurface measurements. He has authored two technical books and developed eight patented technologies for electromagnetic sensing and machine-learning implementations.

Weichang Li leads the Machine Learning Group at Aramco Americas' Houston Research Center, which he joined in January 2015. Weichang obtained his PhD degree in electrical and oceanographic engineering in 2006, and MS (dual) in EECS and OE in 2002, all from MIT. From 2006 to 2008, he was with Woods Hole Oceanographic Institution (WHOI) under an ONR postdoctoral fellowship. From 2008 to 2015, Weichang was with ExxonMobil Corporate Strategic Research Lab where he led the machine-learning group from 2011 to 2014. His current research focus is on machine learning, statistical signal processing algorithm research, and applications in geophysics, geosciences, and petroleum engineering problems.

Michael Ashby is currently working as a contract petrophysicist 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 Schlumberger. 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 for the SPWLA PDDA SIG as the Chairman (2020–2021) and Vice-Chairman (2019).

Wen Pan is a PhD student at the Petroleum Engineering and Geosystems Department of the University of Texas at Austin, conducting research on machine learning, data analytics, geostatistics, well-log interpretation, reservoir modeling, history matching, and production optimization.

Tianqi Deng is a PhD student at the Petroleum Engineering and Geosystems Department of the University of Texas at Austin, specializing in Bayesian inference-based well-log interpretation.

Honggeun Jo is a PhD student at the Petroleum Engineering and Geosystems Department of the University of Texas at Austin, focusing on reservoir modeling, history matching, machine learning, and data analytics.

Javier E. Santos is a PhD student at the Petroleum Engineering and Geosystems Department of the University of Texas at Austin, studying digital rocks, machine learning, and data analytics.

Lei Fu is a data scientist working for Aramco Americas. Lei has 5 years of work experience in artificial intelligence. He received his PhD degree in earth science from Rice University in 2016. Lei loves taking on data challenges and developing AI technology solutions.

Chengran Wang graduated from the University of Houston in 2015 with a master's degree in industrial engineering. Specialized in implementing statistical analysis in business operations, Chengran provides management and supply chain solutions for government and corporations in the US and Canada. She has a great passion for data and technology.

Arkhat Kalbekov is a graduate student at the Petroleum Engineering Department of the Colorado School of Mines. His areas of interest are rock physics (poroelastic behavior of carbonates) and data science. He has 5+ years of experience in the oil and gas sector in general and 3 years as a well-logging engineer with Halliburton in Saudi Arabia in particular, where he performed such geophysical services as quad combo, borehole imaging, NMR logging, and dielectric logging, both onshore and offshore.

Valeria Suarez is a problem-solving engineer with over 3 years of experience in data analysis, petrophysics, and reservoir characterization. Her academic projects are related to fields in the Gulf of Mexico and New Mexico. Previous work in the oil and gas industry focused on managing public and private data to identify trends, optimize well performance, and increase revenue in Texas and Louisiana fields.

Epo Prasetya Kusumah is a lecturer for the Department of Geology, Universitas Pertamina. He teaches a sedimentology-related course and has participated in several company-supported research projects in sedimentology and computer-related geology.

Mohammad Aviandito is a data analyst with domain knowledge in consumer tech and energy. He occasionally writes about data analytics in Bahasa Indonesia for aviandito.medium.com. Prior to his current role, Avi worked for 3 years as a petrophysicist for a major oilfield services company.

Yogi Pamadya is a data analyst and geoscientist who loves to harness the power of Python for geological data analysis. He occasionally writes about geology and data analytics in *Medium*. Yogi realizes that combining an industrial approach of data analytics for classic geological problems can be fun and insightful.

Hossein Izadi is a PhD student at the University of Alberta focusing his research and studies in the areas of mining and petroleum engineering. He is currently the President of the Society of Petroleum Engineers (SPE) Student Chapter at the University of Alberta. He has published 11 journal papers, developed six research-based software, and asserted one patent. He also was an invited speaker about ML and MV applications in geosciences at Zhejiang University–China.