Supplementary File

Predicting brittleness indices of prospective shale formations from sparse well-log suites assisted by derivative and volatility attributes

David A. Wood
DWA Energy Limited, Lincoln, United Kingdom,
ORCID: 0000-0003-3202-4069
Email: dw@dwasolutions.com

This file includes material that complements and expands upon the main article:

Wood, D. A. Predicting brittleness indices of prospective shale formations from sparse well-log suites assisted by derivative and volatility attributes. Advances in Geo-Energy Research, 2022, 6(4): 334-346.

The link to this file is: https://doi.org/10.46690/ager.2022.04.08

Contents

Section S1. Well log curves and their statistical analysis

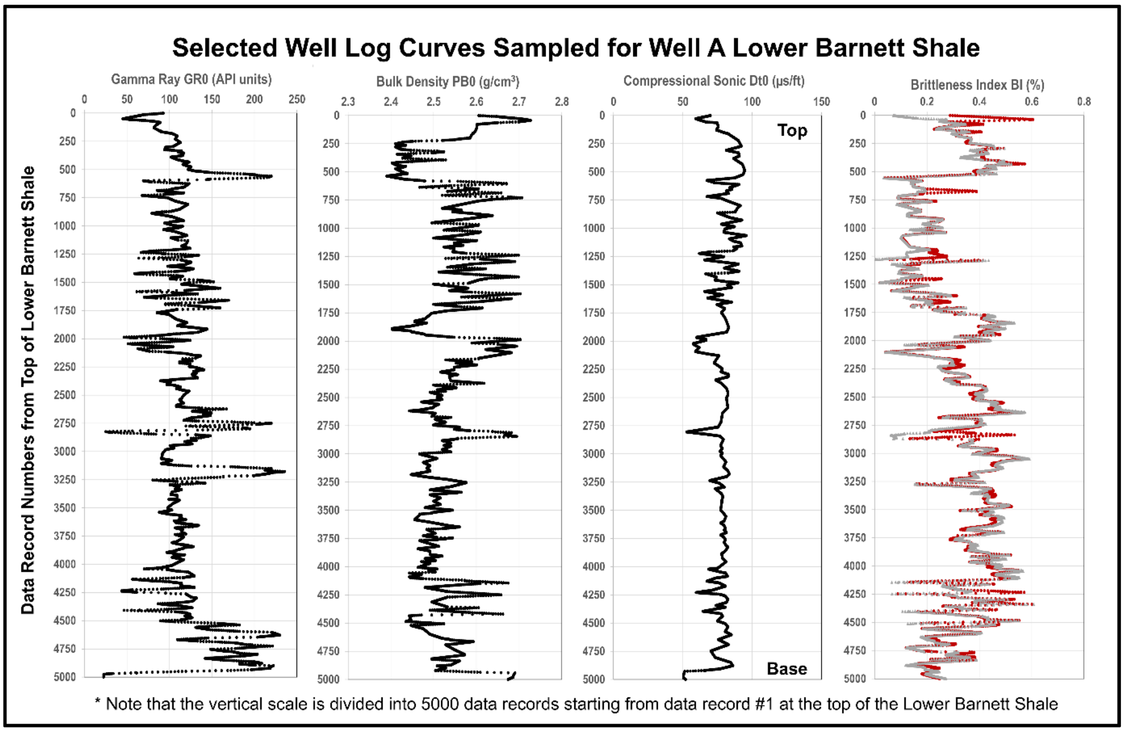


Figure S1. GR0, PB0, DT0 and BI curves for the 470-ft thick LBS formation sampled at Well A. The red BI curve is calculated using the Wang and Gale (2009) method and the grey BI curve is calculated using the Jarvie et al. (2007) method.

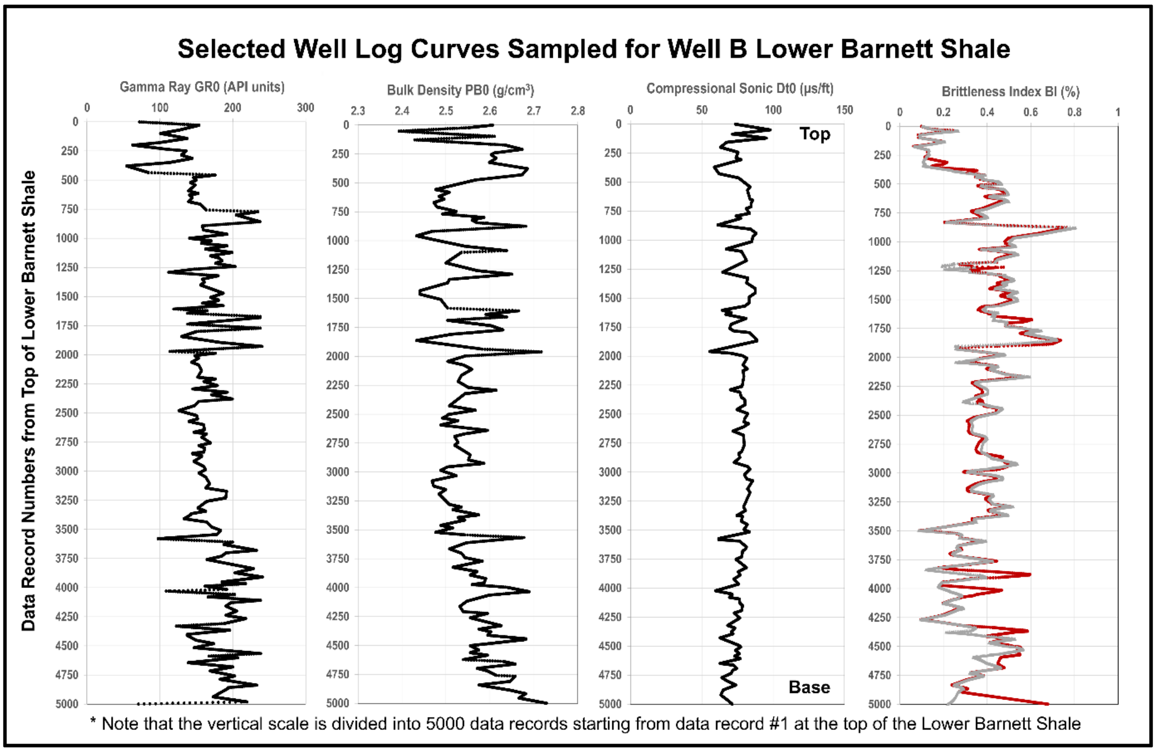
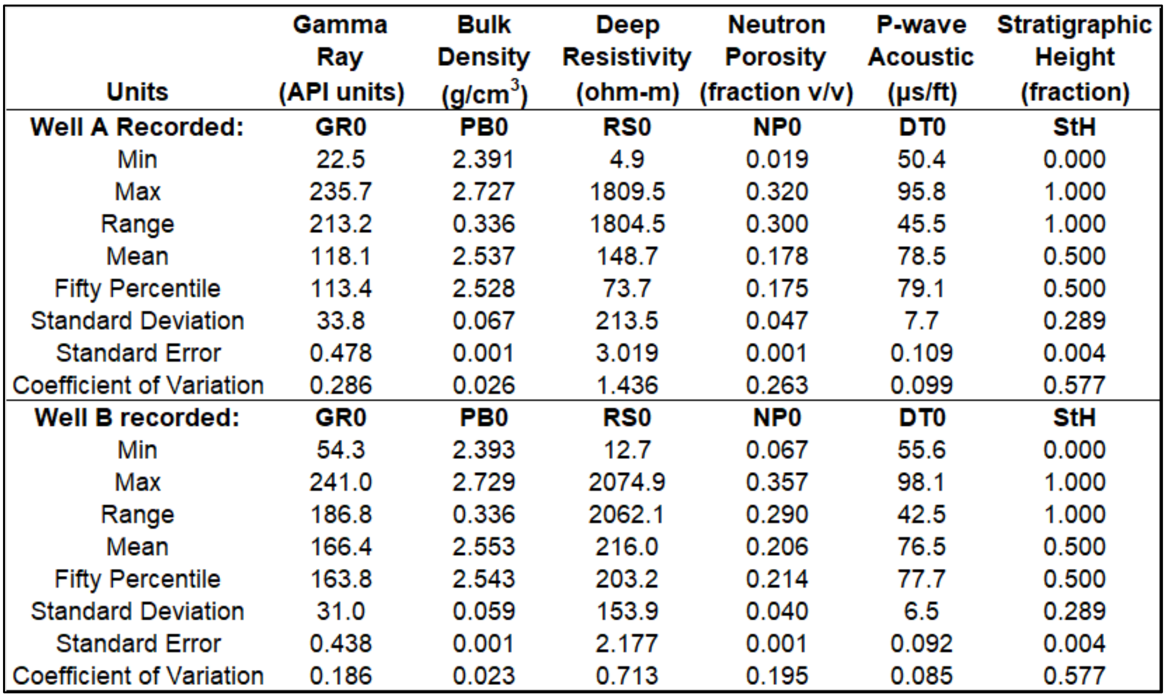


Figure S2. GR0, PB0, DT0 and BI curves for the 300-ft thick LBS formation sampled at Well B. The red BI curve is calculated using the Wang and Gale (2009) method and the grey BI curve is calculated using the Jarvie et al. (2007) method.

Table S1. Statistical summary of Lower Barnett Shale recorded well log variables in Well A and Well B (data from Verma et al., 2016). P-wave refers to the compressional sonic log.



Units	Gamma Ray (API units)	Bulk Density (g/cm^3)	Deep Resistivity (ohm-m)	Neutron Porosity (fraction v/v)	P-wave Acoustic (μs/ft)	Stratigraphic Height (fraction)
Well A Recorded:	**GR0**	**PB0**	**RS0**	**NP0**	**DT0**	**StH**
Min	22.5	2.391	4.9	0.019	50.4	0.000
Max	235.7	2.727	1809.5	0.320	95.8	1.000
Range	213.2	0.336	1804.5	0.300	45.5	1.000
Mean	118.1	2.537	148.7	0.178	78.5	0.500
Fifty Percentile	113.4	2.528	73.7	0.175	79.1	0.500
Standard Deviation	33.8	0.067	213.5	0.047	7.7	0.289
Standard Error	0.478	0.001	3.019	0.001	0.109	0.004
Coefficient of Variation	0.286	0.026	1.436	0.263	0.099	0.577
Well B recorded:	**GR0**	**PB0**	**RS0**	**NP0**	**DT0**	**StH**
Min	54.3	2.393	12.7	0.067	55.6	0.000
Max	241.0	2.729	2074.9	0.357	98.1	1.000
Range	186.8	0.336	2062.1	0.290	42.5	1.000
Mean	166.4	2.553	216.0	0.206	76.5	0.500
Fifty Percentile	163.8	2.543	203.2	0.214	77.7	0.500
Standard Deviation	31.0	0.059	153.9	0.040	6.5	0.289
Standard Error	0.438	0.001	2.177	0.001	0.092	0.004
Coefficient of Variation	0.186	0.023	0.713	0.195	0.085	0.577

Section S2. Hyperparameter values applied to multi-linear regressions and machine learning models utilized

LR: no hyperparameters requiring adjustment.

ElasticNet: alpha =0.0001; L1 ratio =0.4.

K- nearest neighbor (KNN): K (number of neighbors considered) = 2; distance measure = Manhattan.

Support Vector Regression (SVR): kernel = radial basis function (RBF); C (penalty parameter of the error term) = 300; gamma (curvature weight of the decision boundary) =20.

Adaptive Boosting (ADA): number of estimators = 500; maximum depth = 50; learning rate = 0.01; loss function =exponential; splitter = best; splitting criterion = mean squared error (mse).

Random Forest (RF): number of estimators = 750; maximum depth = 50; splitting criterion = mse.

Extreme Gradient Boosting (XGB): number of estimators = 1000; maximum depth = 15; eta = 0.03; columns sampled per tree =0.9; subsample = 0.6.

Section S3. Data normalization

Each well log and attribute is normalized such that its values are distributed on a scale of -1 to +1. This is necessary precaution to avoid scaling biases affecting the prediction models and is achieved by applying Eq. A1 to each variable.

$$Normx_i^n = 2 * \left(\frac{x_i^n - xmin^n}{xmax^n - xmin^n}\right) - 1 \qquad \text{(A1)}$$

where $Normx_i^n$ is the normalized value of the i^{th} data-record relating to the n^{th} variable distribution, $\mathrm{x_i^n}$ is the actual recorded /calculated well-log or attribute value, $xmin^n$ and $xmax^n$ are the minimum and maximum recorded/calculated values associated with the n^{th} variable, respectively.

Section S4. Statistical measures of prediction performance

The statistical error-assessment metrics used to monitor and compare BI prediction performance are expressed in Eqs. (A2) to (A4).

Mean Absolute Error (MAE)

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|rDV_i - pDV_i| \qquad \text{(A2)}$$

where rDV_i is the recorded BI value, i.e., the dependent variable (*DV)*, and pDV_i is the predicted value of i^{th} data record, and *m* is the number of data records in the validation subset being considered.

Root Mean Squared Error (RMSE)

$$RMSE = \left[\frac{1}{m}\sum_{i=1}^{m}((rDV_i) - (pDV_i))^2\right]^{\frac{1}{2}} \qquad \text{(A3)}$$

For the *DV*s considered, MAE and RMSE values are expressed in BI units relative to the range 0 to 1. Hence, an MAE or RMSE value of 0.01 represents 1% of that range.

Coefficient of Determination (R^2)

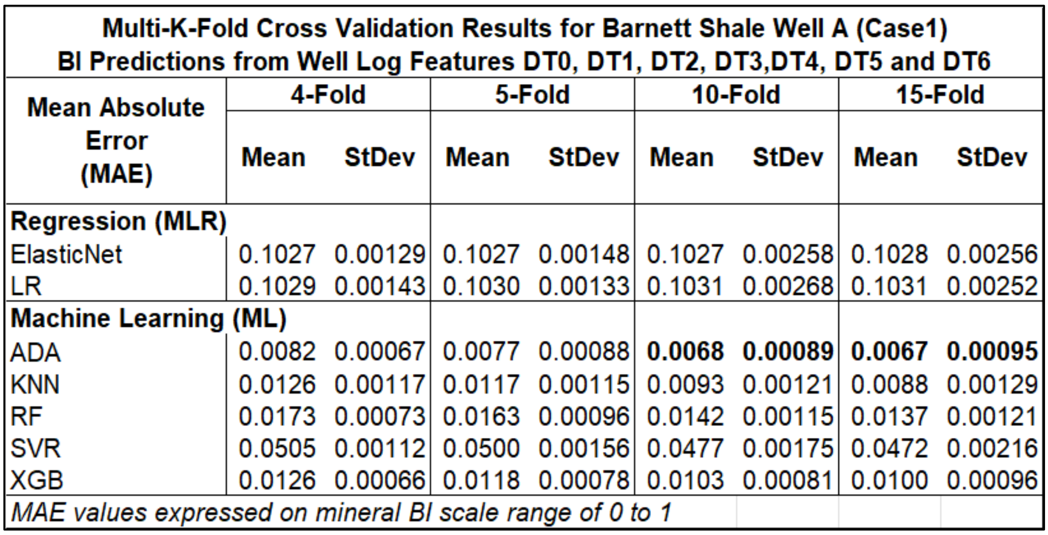
$$R^2 = \left\{\frac{\sum_{i=1}^{m}(rDV_i - \overline{rDV})(pDV_i - \overline{pDV})}{\sqrt{\sum_{i=1}^{m}(rDV_i - \overline{rDV})^2}\sqrt{\sum_{i=1}^{m}(pDV_i - \overline{pDV})^2}}\right\}^2 \qquad \text{(A4)}$$

where $\overline{rDV}$ and $\overline{pDV}$ are mean values of distributions *mDV* and *pDV*, respectively. The R^2 value, the square of the Pearson correlation coefficient, varies between 0 to 1.

Section S5. Case 1 Results involving the DT well log and its attributes applied to Well A

Table S2 displays the multi-fold cross-validation Case 1 results for each of the MLR/ML models applied to Well A.

Table S2. Multi-K-fold analysis results for MLR and ML models applied to the Case1 Well A dataset.



Multi-K-Fold Cross Validation Results for Barnett Shale Well A (Case1)
BI Predictions from Well Log Features DT0, DT1, DT2, DT3,DT4, DT5 and DT6

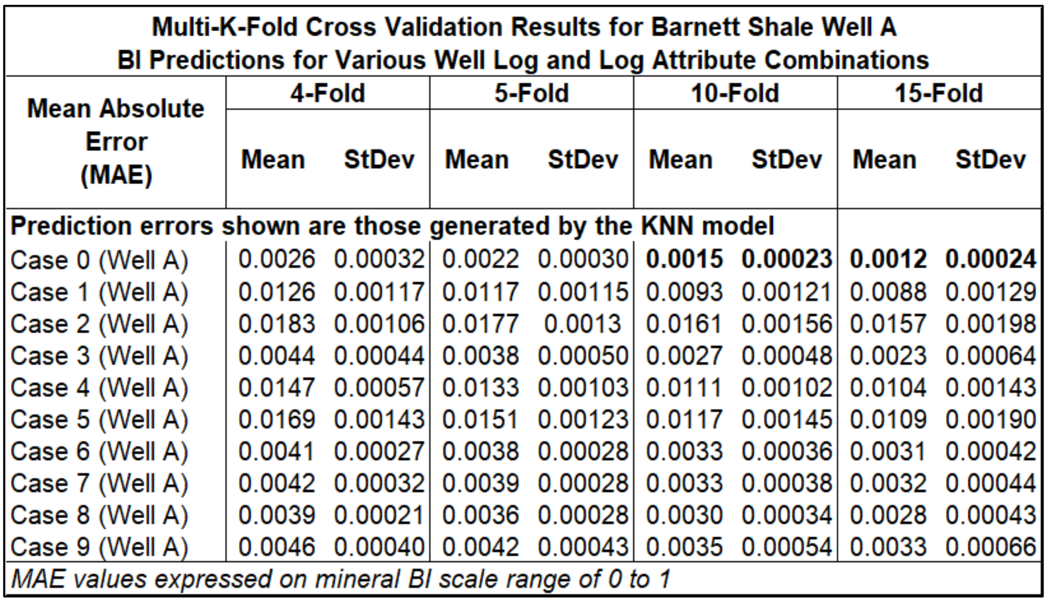
Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Regression (MLR)								
ElasticNet	0.1027	0.00129	0.1027	0.00148	0.1027	0.00258	0.1028	0.00256
LR	0.1029	0.00143	0.1030	0.00133	0.1031	0.00268	0.1031	0.00252
Machine Learning (ML)								
ADA	0.0082	0.00067	0.0077	0.00088	**0.0068**	**0.00089**	**0.0067**	**0.00095**
KNN	0.0126	0.00117	0.0117	0.00115	0.0093	0.00121	0.0088	0.00129
RF	0.0173	0.00073	0.0163	0.00096	0.0142	0.00115	0.0137	0.00121
SVR	0.0505	0.00112	0.0500	0.00156	0.0477	0.00175	0.0472	0.00216
XGB	0.0126	0.00066	0.0118	0.00078	0.0103	0.00081	0.0100	0.00096

MAE values expressed on mineral BI scale range of 0 to 1

Section S6. Multi-K-fold cross validation analysis for all cases relating to Well B

The multi-K-fold cross validation analysis for all ten cases modelled separately for well A and Well B with the KNN model are displayed in Tables S3 and S4, respectively. The benchmark Case 0 generates the lowest BI prediction error of the models considered for Wells A and B. However, Cases 6 to 9, involving fewer recorded well logs, also generate BI predictions with very low errors for Wells A and B.

Table S3. Multi-K-fold analysis results for ten cases of distinct well-log and attribute combinations assessed for Well A with the KNN prediction model.

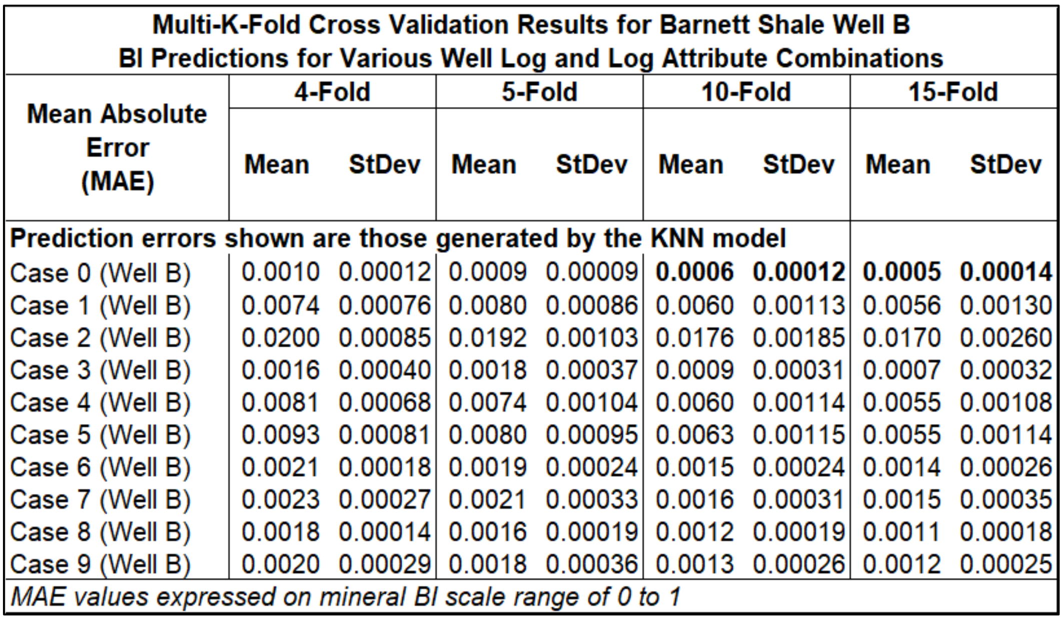


Multi-K-Fold Cross Validation Results for Barnett Shale Well A
BI Predictions for Various Well Log and Log Attribute Combinations

Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Prediction errors shown are those generated by the KNN model								
Case 0 (Well A)	0.0026	0.00032	0.0022	0.00030	**0.0015**	**0.00023**	**0.0012**	**0.00024**
Case 1 (Well A)	0.0126	0.00117	0.0117	0.00115	0.0093	0.00121	0.0088	0.00129
Case 2 (Well A)	0.0183	0.00106	0.0177	0.0013	0.0161	0.00156	0.0157	0.00198
Case 3 (Well A)	0.0044	0.00044	0.0038	0.00050	0.0027	0.00048	0.0023	0.00064
Case 4 (Well A)	0.0147	0.00057	0.0133	0.00103	0.0111	0.00102	0.0104	0.00143
Case 5 (Well A)	0.0169	0.00143	0.0151	0.00123	0.0117	0.00145	0.0109	0.00190
Case 6 (Well A)	0.0041	0.00027	0.0038	0.00028	0.0033	0.00036	0.0031	0.00042
Case 7 (Well A)	0.0042	0.00032	0.0039	0.00028	0.0033	0.00038	0.0032	0.00044
Case 8 (Well A)	0.0039	0.00021	0.0036	0.00028	0.0030	0.00034	0.0028	0.00043
Case 9 (Well A)	0.0046	0.00040	0.0042	0.00043	0.0035	0.00054	0.0033	0.00066

MAE values expressed on mineral BI scale range of 0 to 1

Table S4. Multi-K-fold analysis results for ten cases of distinct well-log and attribute combinations assessed for Well B with the KNN prediction model.



Multi-K-Fold Cross Validation Results for Barnett Shale Well B
BI Predictions for Various Well Log and Log Attribute Combinations

Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Prediction errors shown are those generated by the KNN model								
Case 0 (Well B)	0.0010	0.00012	0.0009	0.00009	**0.0006**	**0.00012**	**0.0005**	**0.00014**
Case 1 (Well B)	0.0074	0.00076	0.0080	0.00086	0.0060	0.00113	0.0056	0.00130
Case 2 (Well B)	0.0200	0.00085	0.0192	0.00103	0.0176	0.00185	0.0170	0.00260
Case 3 (Well B)	0.0016	0.00040	0.0018	0.00037	0.0009	0.00031	0.0007	0.00032
Case 4 (Well B)	0.0081	0.00068	0.0074	0.00104	0.0060	0.00114	0.0055	0.00108
Case 5 (Well B)	0.0093	0.00081	0.0080	0.00095	0.0063	0.00115	0.0055	0.00114
Case 6 (Well B)	0.0021	0.00018	0.0019	0.00024	0.0015	0.00024	0.0014	0.00026
Case 7 (Well B)	0.0023	0.00027	0.0021	0.00033	0.0016	0.00031	0.0015	0.00035
Case 8 (Well B)	0.0018	0.00014	0.0016	0.00019	0.0012	0.00019	0.0011	0.00018
Case 9 (Well B)	0.0020	0.00029	0.0018	0.00036	0.0013	0.00026	0.0012	0.00025
MAE values expressed on mineral BI scale range of 0 to 1								

Section S7. Relative influence of recorded well logs for Case 3

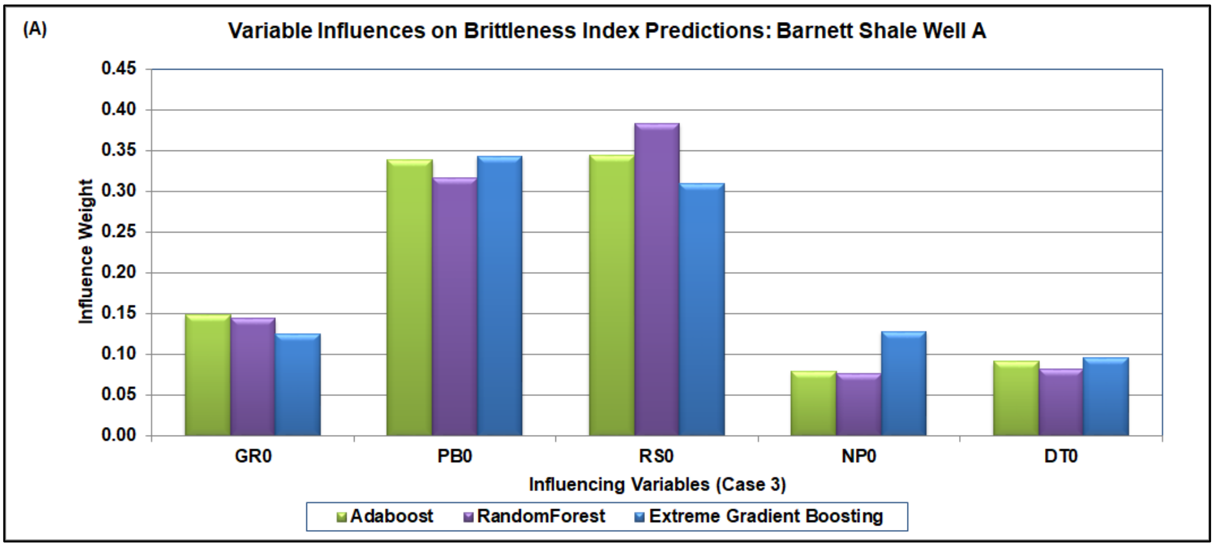
The relative influence analysis for Case 3 (GR0, PB0, RS0,NP0, DT0) reveals that PB0 and RS0 dominate the solutions for Well A, whereas RS0, NP0 and GR0 exert most influence on the Well B solutions (Figure S3). For Well A the relative order of influence is:

RS0 ≈PB0 >GR0 >DT0 ≈NP0.

For Well B the relative order of influence is:

NP0 ≈RS0 >GR0 >PB0 ≈DT0.

The prediction performance of Case 3 is only slightly inferior to that of benchmark Case 0 that additionally incorporates variable StH.



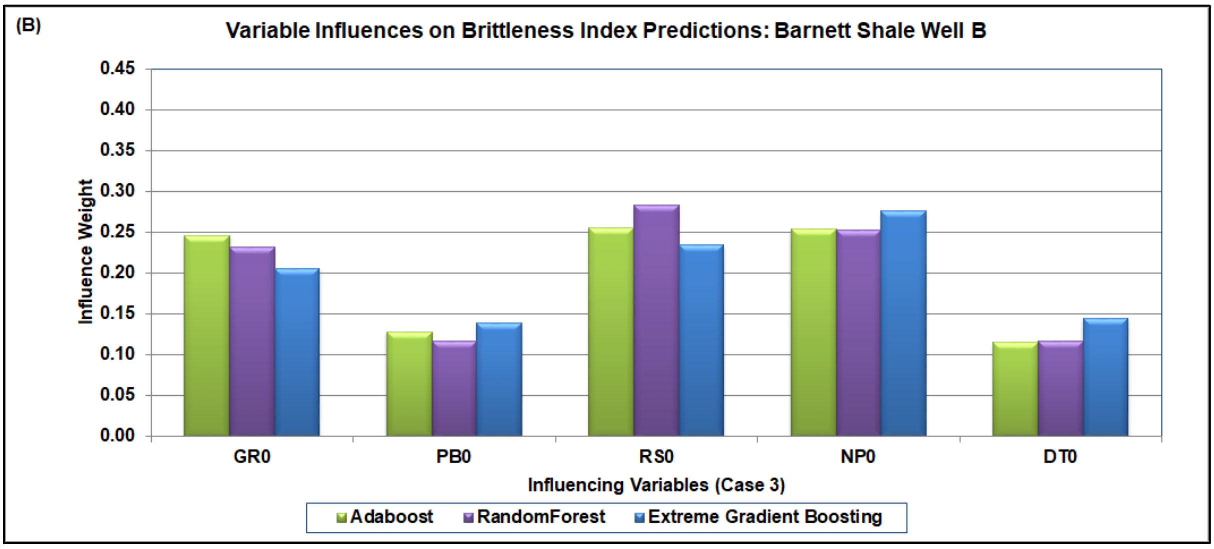
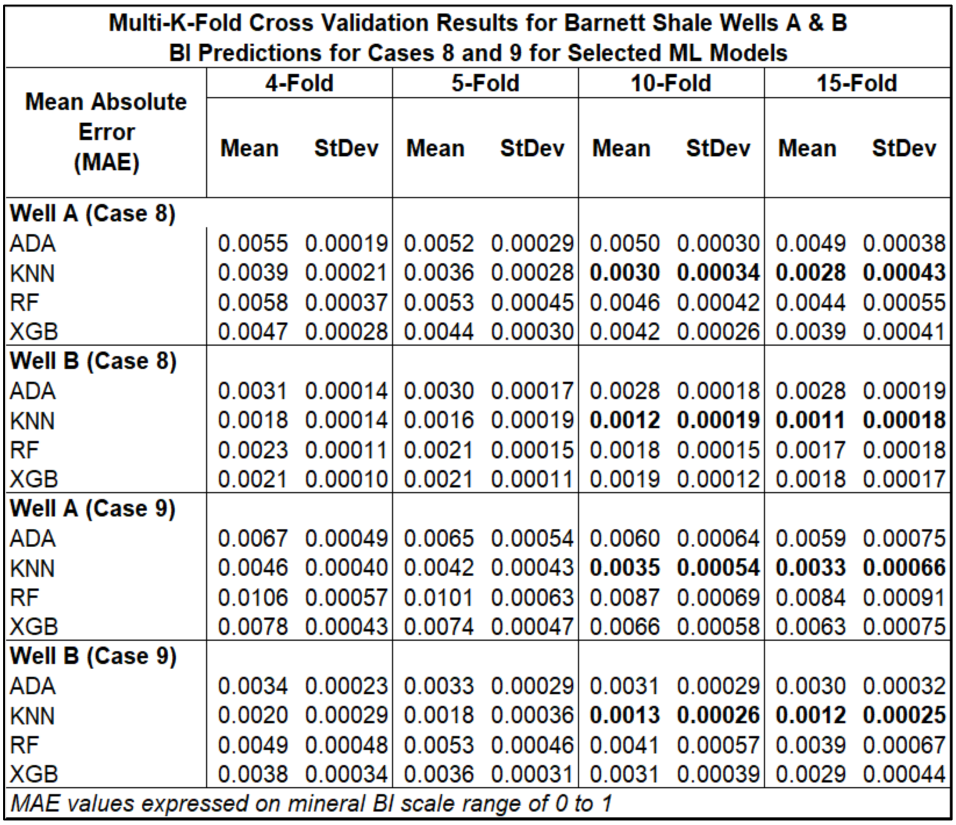


Figure S3. Relative importance of all recorded well logs to the tree-ensemble model solutions applied to 5-variable Case 3 for: (A) Well A; and (B) Well B.

Section S8. BI prediction results and relative feature influences for Cases 8 and 9

Table S5 displays multi-K-fold analysis for KNN and the three tree-ensemble models applied to Cases 8 and 9. All models assessed provide accuracy that rivals that achieved by Case 3, with KNN slightly outperforming the other models for all four K-folds considered.



Multi-K-Fold Cross Validation Results for Barnett Shale Wells A & B
BI Predictions for Cases 8 and 9 for Selected ML Models

Mean Absolute Error (MAE)	4-Fold		5-Fold		10-Fold		15-Fold	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Well A (Case 8)								
ADA	0.0055	0.00019	0.0052	0.00029	0.0050	0.00030	0.0049	0.00038
KNN	0.0039	0.00021	0.0036	0.00028	**0.0030**	**0.00034**	**0.0028**	**0.00043**
RF	0.0058	0.00037	0.0053	0.00045	0.0046	0.00042	0.0044	0.00055
XGB	0.0047	0.00028	0.0044	0.00030	0.0042	0.00026	0.0039	0.00041
Well B (Case 8)								
ADA	0.0031	0.00014	0.0030	0.00017	0.0028	0.00018	0.0028	0.00019
KNN	0.0018	0.00014	0.0016	0.00019	**0.0012**	**0.00019**	**0.0011**	**0.00018**
RF	0.0023	0.00011	0.0021	0.00015	0.0018	0.00015	0.0017	0.00018
XGB	0.0021	0.00010	0.0021	0.00011	0.0019	0.00012	0.0018	0.00017
Well A (Case 9)								
ADA	0.0067	0.00049	0.0065	0.00054	0.0060	0.00064	0.0059	0.00075
KNN	0.0046	0.00040	0.0042	0.00043	**0.0035**	**0.00054**	**0.0033**	**0.00066**
RF	0.0106	0.00057	0.0101	0.00063	0.0087	0.00069	0.0084	0.00091
XGB	0.0078	0.00043	0.0074	0.00047	0.0066	0.00058	0.0063	0.00075
Well B (Case 9)								
ADA	0.0034	0.00023	0.0033	0.00029	0.0031	0.00029	0.0030	0.00032
KNN	0.0020	0.00029	0.0018	0.00036	**0.0013**	**0.00026**	**0.0012**	**0.00025**
RF	0.0049	0.00048	0.0053	0.00046	0.0041	0.00057	0.0039	0.00067
XGB	0.0038	0.00034	0.0036	0.00031	0.0031	0.00039	0.0029	0.00044

MAE values expressed on mineral BI scale range of 0 to 1

Table S5. Multi-K-fold analysis results of feature-selected Cases 8 and 9 to predict BI for wells A and B applying KNN and three tree-ensemble prediction models.

Variables StH and PB0 exert the dominant influences (weights ~0.35) for the Case 8 Well A model solutions (Figure S4A), with StH being substantially more influential than other variables for Case 8 Well B (Figure S4B). Variables GR1, PB1 and DT1 exert more influence in the XGB model than the ADA and RF models in Case 8 solutions for both wells.

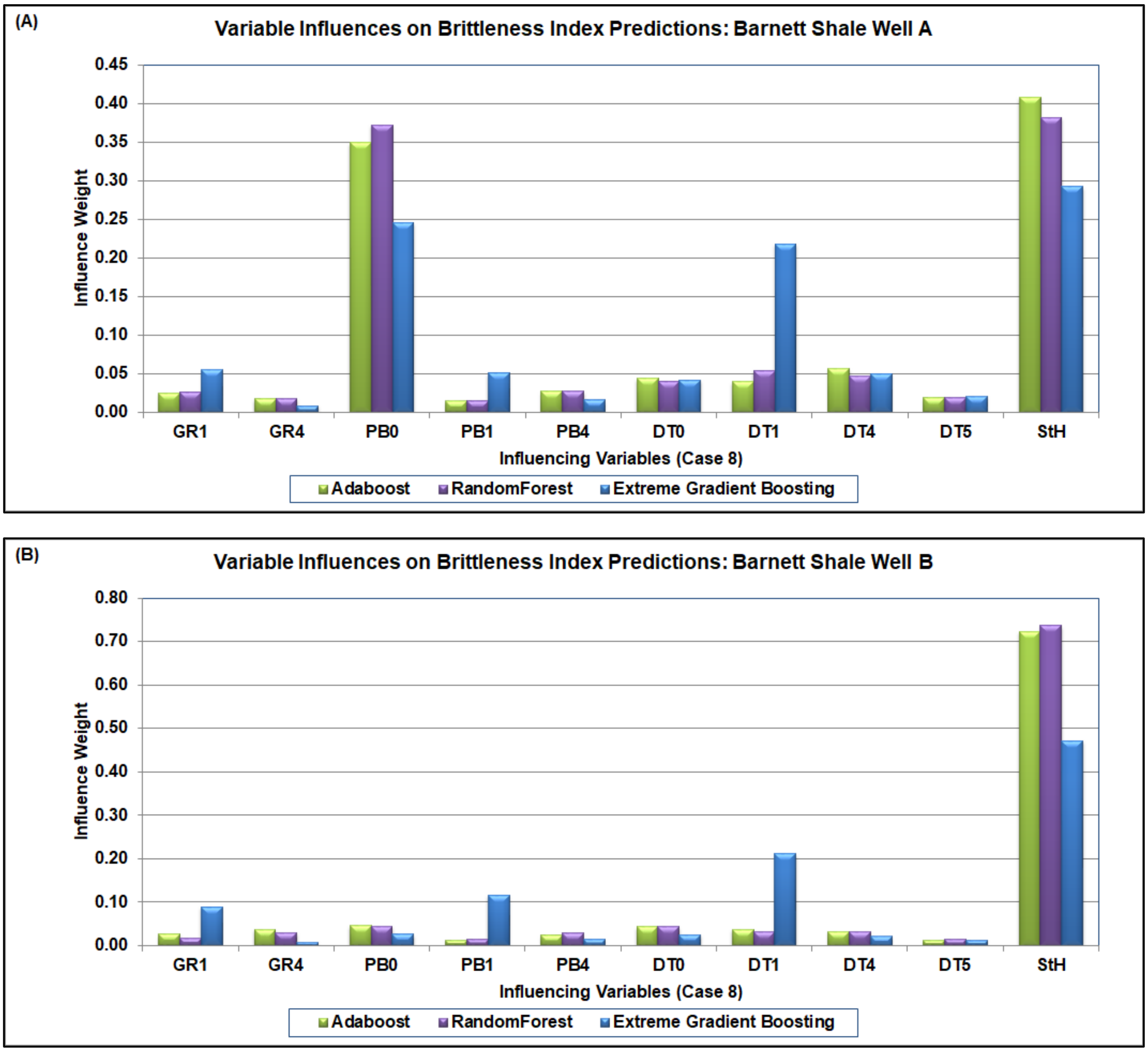
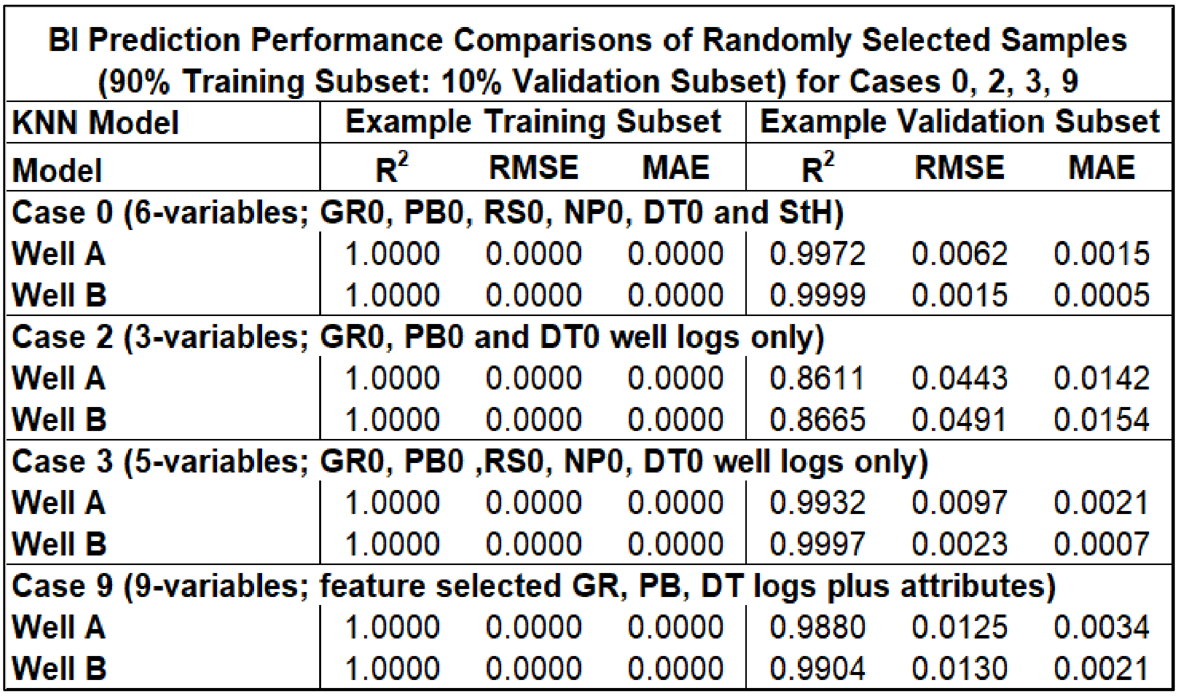


Figure S4. Relative importance of 10 feature-selected variables to the tree-ensemble model solutions applied to Case 8 for: (A) Well A; and (B) Well B.

Section S9. Random subset prediction performances for Cases 0, 2, 3 and 9

The prediction accuracy for Case 9 involving feature-selected attributes is substantially improved versus Case 3. Comparisons of the BI prediction performances, in terms of MAE, RMSE and R^2, of the KNN models for example validation subsets relating to Cases 0, 2, 3, and 9 are shown in Table S6. The feature-selected Case 9 solution, based on only the GR, PB and DT recorded well logs plus selected attributes delivers only slightly inferior BI prediction results to those involving 5 recorded well logs.

Table S6. Examples of randomly selected training and validation subset BI prediction performances for the KNN model for Cases 0, 2, 3, and 9. The data records for these subset examples for Cases 2 and 9 are displayed in Figure 9 of the main article.

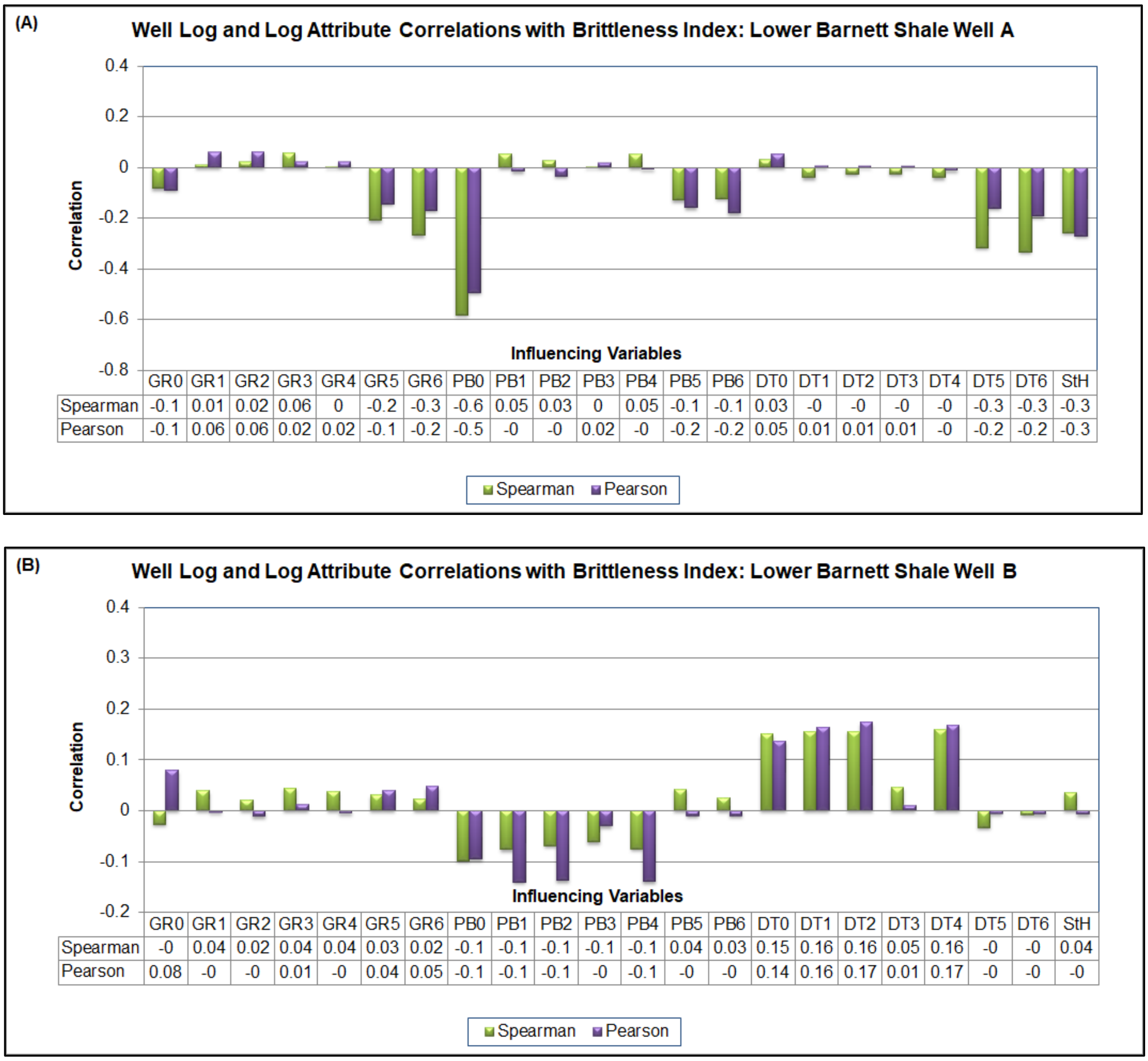


BI Prediction Performance Comparisons of Randomly Selected Samples (90% Training Subset: 10% Validation Subset) for Cases 0, 2, 3, 9

KNN Model	Example Training Subset			Example Validation Subset		
Model	R^2	RMSE	MAE	R^2	RMSE	MAE
Case 0 (6-variables; GR0, PB0, RS0, NP0, DT0 and StH)						
Well A	1.0000	0.0000	0.0000	0.9972	0.0062	0.0015
Well B	1.0000	0.0000	0.0000	0.9999	0.0015	0.0005
Case 2 (3-variables; GR0, PB0 and DT0 well logs only)						
Well A	1.0000	0.0000	0.0000	0.8611	0.0443	0.0142
Well B	1.0000	0.0000	0.0000	0.8665	0.0491	0.0154
Case 3 (5-variables; GR0, PB0 ,RS0, NP0, DT0 well logs only)						
Well A	1.0000	0.0000	0.0000	0.9932	0.0097	0.0021
Well B	1.0000	0.0000	0.0000	0.9997	0.0023	0.0007
Case 9 (9-variables; feature selected GR, PB, DT logs plus attributes)						
Well A	1.0000	0.0000	0.0000	0.9880	0.0125	0.0034
Well B	1.0000	0.0000	0.0000	0.9904	0.0130	0.0021

Section S10. Correlation coefficients between Well Log attributes and BI

Figure S5 displays Pearson (*R*) and Spearman (*P*) correlation coefficients between the well log attributes considered and BI for Wells A and B, to highlight this point. The correlations between the well-log attributes and BI is quite distinct for the two wells considered. These differences have undoubtedly affected the influence of the attributes on the prediction model solutions, as revealed by comparing Figures 5 and 6 (main text) with Figure S5. For Well A (Figure S5A), there are substantial differences between *P* and *R* values for most of the well logs and well-log attributes versus BI. This suggests that few, if any, of the attribute relationships with BI can be considered as even approximately parametric. This is also the case for Well B (Figure S5B), but less so, as for PB0, DT0 and most of the DT attributes *P* and *R* values are in closer agreement than for Well A.



	GR0	GR1	GR2	GR3	GR4	GR5	GR6	PB0	PB1	PB2	PB3	PB4	PB5	PB6	DT0	DT1	DT2	DT3	DT4	DT5	DT6	StH
Spearman	-0.1	0.01	0.02	0.06	0	-0.2	-0.3	-0.6	0.05	0.03	0	0.05	-0.1	-0.1	0.03	-0	-0	-0	-0	-0.3	-0.3	-0.3
Pearson	-0.1	0.06	0.06	0.02	0.02	-0.1	-0.2	-0.5	-0	-0	0.02	-0	-0.2	-0.2	0.05	0.01	0.01	0.01	-0	-0.2	-0.2	-0.3

	GR0	GR1	GR2	GR3	GR4	GR5	GR6	PB0	PB1	PB2	PB3	PB4	PB5	PB6	DT0	DT1	DT2	DT3	DT4	DT5	DT6	StH
Spearman	-0	0.04	0.02	0.04	0.04	0.03	0.02	-0.1	-0.1	-0.1	-0.1	-0.1	0.04	0.03	0.15	0.16	0.16	0.05	0.16	-0	-0	0.04
Pearson	0.08	-0	-0	0.01	-0	0.04	0.05	-0.1	-0.1	-0.1	-0	-0.1	-0	-0	0.14	0.16	0.17	0.01	0.17	-0	-0	-0

Figure S5. Pearson and Spearman correlation coefficients for well logs and attributes for GR, PB and DT with the calculated Wang and Gale BI index for: (A) Well A; and, (B) Well B.