Well-log attributes assist in the determination of reservoir formation tops in wells with sparse well-log data

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

Supplementary File

-- David A. Wood

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

Wood, D. A. Well-log attributes assist in the determination of reservoir formation tops in wells with sparse well-log data. Advances in Geo-Energy Research, 2023, 8 (1), 45-60. https://doi.org/10.46690/ager.2023.04.05.

The link to this file is: https://doi.org/10.46690/ager.2023.04.05

Contents

Section S1. Well-log curves evaluated for Wells A and B

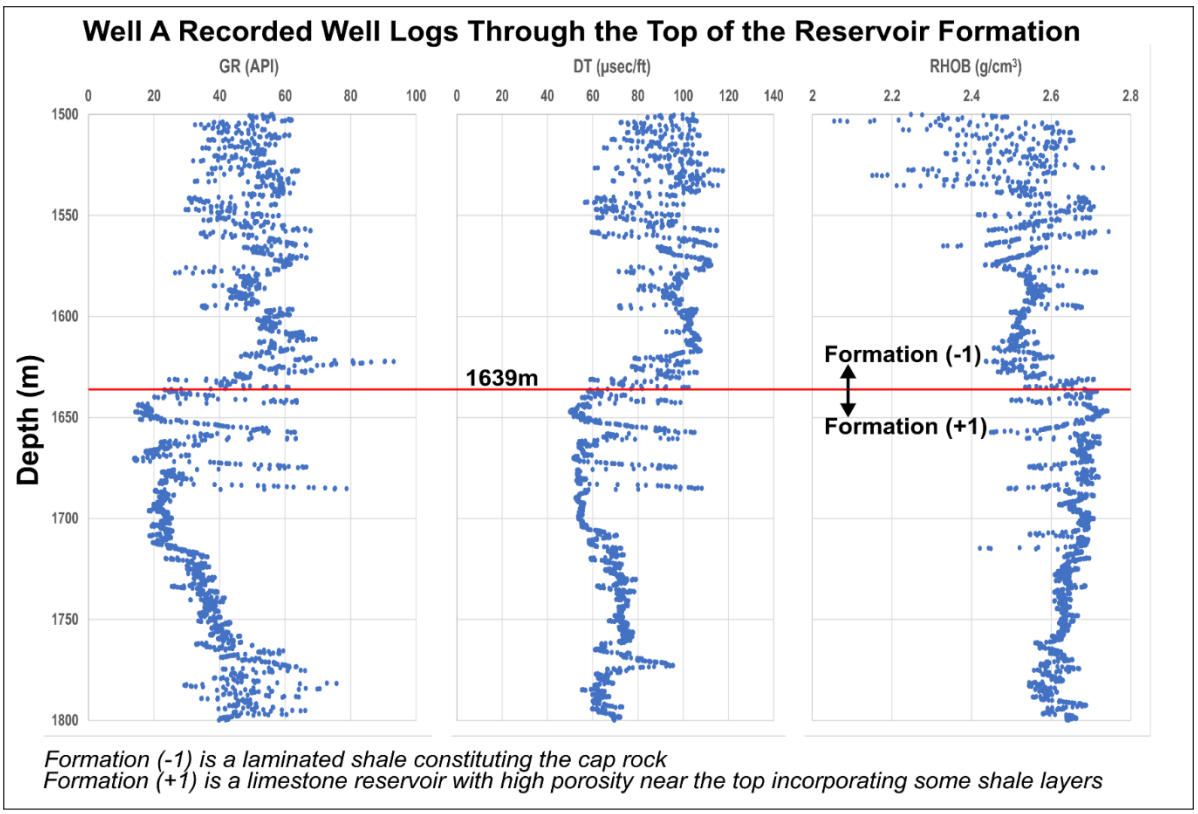

Figure S1. Recorded well logs (one recording each 0.5ft (or 0.1524m) for Well A displayed over interval straddling the cap rock (Formation (-1)) and the top of the reservoir (Formation (+1). 1969 recorded samples used for analysis (918 in Formation (-1) and 1051 in Formation (+1).

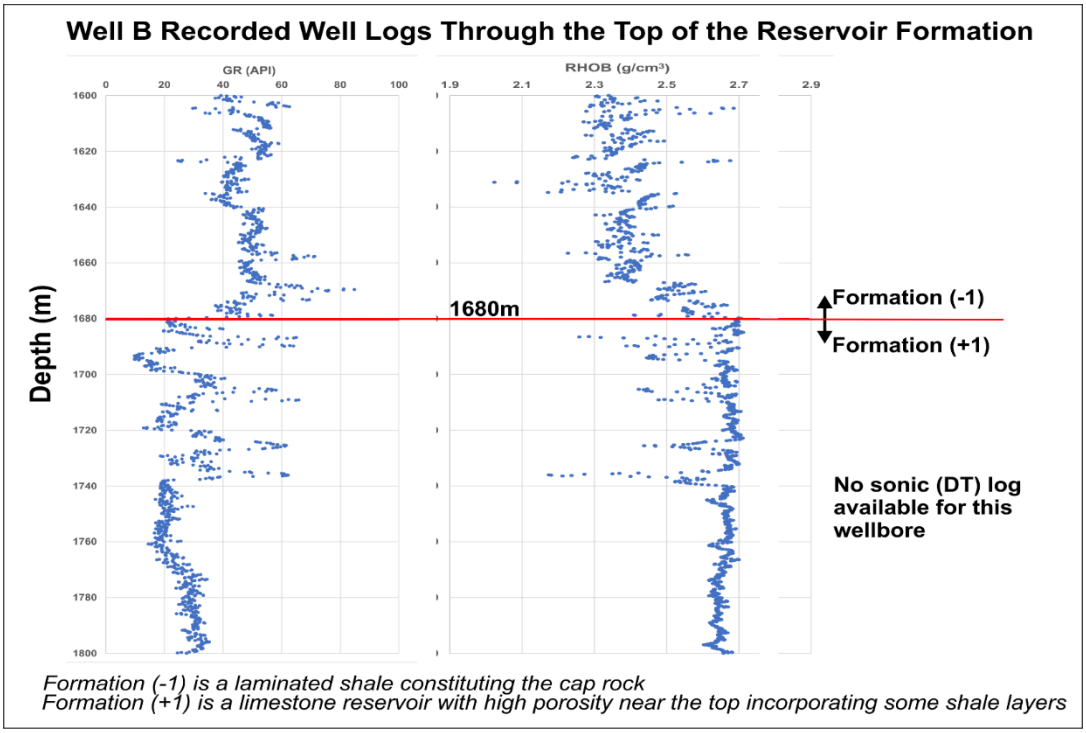

Figure S2. Recorded well logs (one recording each 0.5ft (or 0.1524m) for Well B displayed over interval straddling the cap rock (Formation (-1)) and the top of the reservoir (Formation (+1). 1314 recorded samples used for analysis (524 in Formation (-1) and 790 in Formation (+1).

Section S2. Formulas for statistical error metrics assessed

By assigning the two formation categories the numerical values of -1 and +1, in addition to accuracy, it is also possible to calculate other numerical statistical error metrics, including mean absolute error (MAE; Eq. (E1)), root mean squared error (RMSE; Eq. (E2)), and coefficient of determination (R^2, Eq. (E3) that provide complementary prediction performance information.

Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \quad \text{(E1)}$$

Root Mean Square Error (MSE):

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}((X_i) - (Y_i))^2\right]^{0.5} \quad \text{(E2)}$$

where X_i = the actual formation category value and Y_i = the predicted formation category value for the i^{th} data record being predicted.

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Xi - Yi)^2}{\sum_{i=1}^{n}(Xi - Xmean)^2} \quad \text{(E3)}$$

Where *Xmean* is the mean of the *X* variable distribution.

Section S3. Optimized feature selections for GR-PB and GR-DT attribute combinations

Table S1. Feature selections for GR-PB recorded logs plus attribute combinations applied to Well A data based on a 0.75 (training): 0.25 (validation) split of data records and applying a KNN prediction model with various optimizers. Only the 8 feature selections (of the 42 optimizer runs performed) that generated the most accurate formation predictions are shown.

Features Selected (Marked by (X)) by KNN with Various Optimizers Using Only GR and PB Logs Plus Attributes								
	Jaya-D	DE-C	PSO-C	CSO-C	SCA-A	Jaya-E	PSO-F	SCA-C
	Fitness Score Increases →							
GR0	X	X	X	X		X	X	X
GR1								
GR2		X	X			X	X	X
GR3								
GR4								
GR5	X	X	X	X	X	X	X	X
GR6	X		X	X	X		X	X
PB0	X	X	X	X	X	X	X	
PB1				X				
PB2	X	X	X	X	X	X	X	
PB3	X							
PB4				X			X	
PB5	X	X	X	X	X	X	X	X
PB6	X	X	X	X	X	X	X	X
Features Selected	8	7	8	8	6	7	9	6
Accuracy (0 to 1)	0.9857	0.9836	0.9836	0.9816	0.9795	0.9795	0.9816	0.9775
Optimizer Population	65	45	60	20	50	60	65	60
Number of Iterations	100	100	100	100	100	100	100	100
Fitness Score	0.0199	0.0212	0.0219	0.0240	0.0246	0.0253	0.0247	0.0266
Execution Time (s)	90.577	63.621	87.143	65.174	37.696	80.71	100.36	62.673

Table 5. Feature selections for GR-DT recorded logs plus attribute combinations applied to Well A data based on a 0.75 (training): 0.25 (validation) split of data records and applying a KNN prediction model with various optimizers. Only the 8 feature selections (of the 42 optimizer runs performed) that generated the most accurate formation predictions are shown.

Features Selected (Marked by (X)) by KNN with Various Optimizers Using Only GR and DT Logs Plus Attributes								
	SCA-F	DE-B	Jaya-A	CSO-G	DE-F	PSO-A	DE-E	PSO-D
	Fitness Score Increases →							
GR0	X	X	X	X	X		X	X
GR1								
GR2		X		X			X	X
GR3	X						X	
GR4								
GR5		X		X	X	X	X	
GR6	X		X				X	X
DT0	X	X	X	X	X	X	X	X
DT1								
DT2	X	X		X	X	X	X	X
DT3								
DT4								
DT5	X	X	X	X		X	X	X
DT6	X	X	X	X	X	X	X	X
Features Selected	7	7	5	7	5	5	9	7
Accuracy (0 to 1)	0.9818	0.9818	0.9898	0.9898	0.9877	0.9877	0.9898	0.9877
Optimizer Population	50	60	40	30	60	100	65	75
Number of Iterations	100	100	100	100	100	100	100	100
Fitness Score	0.0131	0.0131	0.0137	0.0151	0.0157	0.0157	0.0166	0.0172
Execution Time (s)	55.731	87.369	45.971	84.813	60.369	99.87	93.256	98.225

Section S4. Accuracy and error metric relationships

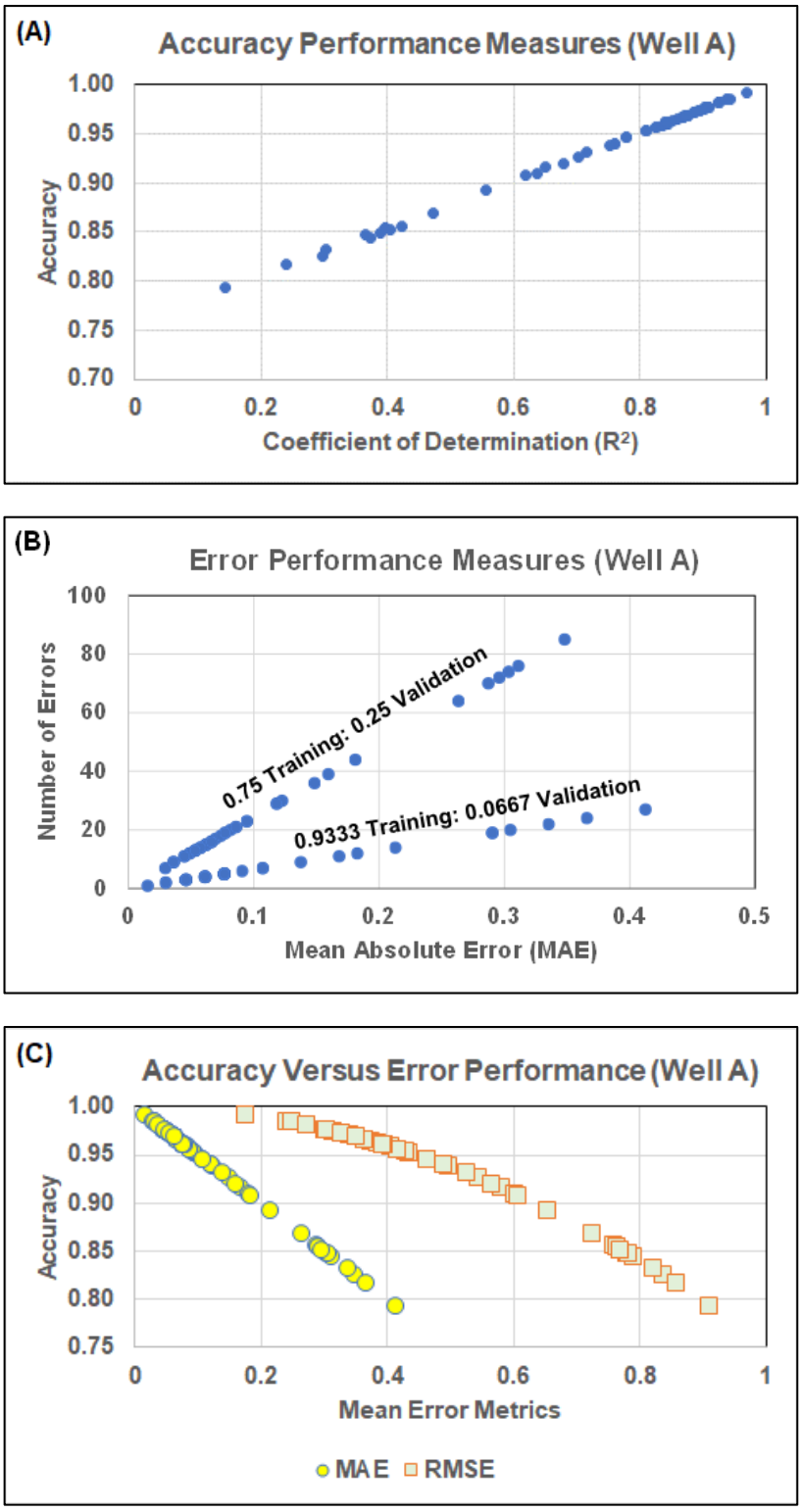

Figure S3. Prediction results of 90 randomly selected validation subsets involving the 15 cases, KNN, SVC and XGB models and 0.25 : 0.75 and 0.9333 : 0.0667 training : validation splits for the Well A dataset.

The advantage of using the numerical category assignments for the two formations of interest (-1 cap rock; +1 reservoir) is apparent in Figure S3 with respect to the alternative error measurements it generates. Figure S3A shows a near perfect linear relationship between accuracy and R^2 among the ninety cases considered, making R^2 a useful proxy for accuracy.

Figure S3B shows near-perfect linear relationships between MAE and the number of errors generated by a prediction model. However, those relationships are dependent on the sample sizes involved; the 0.25 training : 0.75 validation splits involve validation subsets generated from the Well A dataset of 488 data records, whereas the 0.9333 training: 0.0667 validation splits involve validation subsets generated from the Well A of 131 data records. As different subset sizes are involved, for certain purposes, it is generally more meaningful to express errors as percentages in addition to recording the absolute error numbers.

Figure S3C displays the near-perfect linear relationship between the MAE and accuracy metrics. Every correct prediction generates an absolute error of zero whereas every incorrect prediction generates an absolute error value of two (+1 minus -1 or vice versa). Although magnitude of the absolute error value is arbitrary it provides a useful error metric that is directly related to accuracy. On the other hand, the RMSE relationship with accuracy (Figure S3C) is non-linear as each prediction error generates a squared-error value of 4 (2^2) the mean of which is then adjusted to its square root. The RMSE scale provides a more sensitive error scale than MAE for evaluations achieving accuracy of >=0.9.

Section S5. Confusion matrices showing prediction distributions for the best models

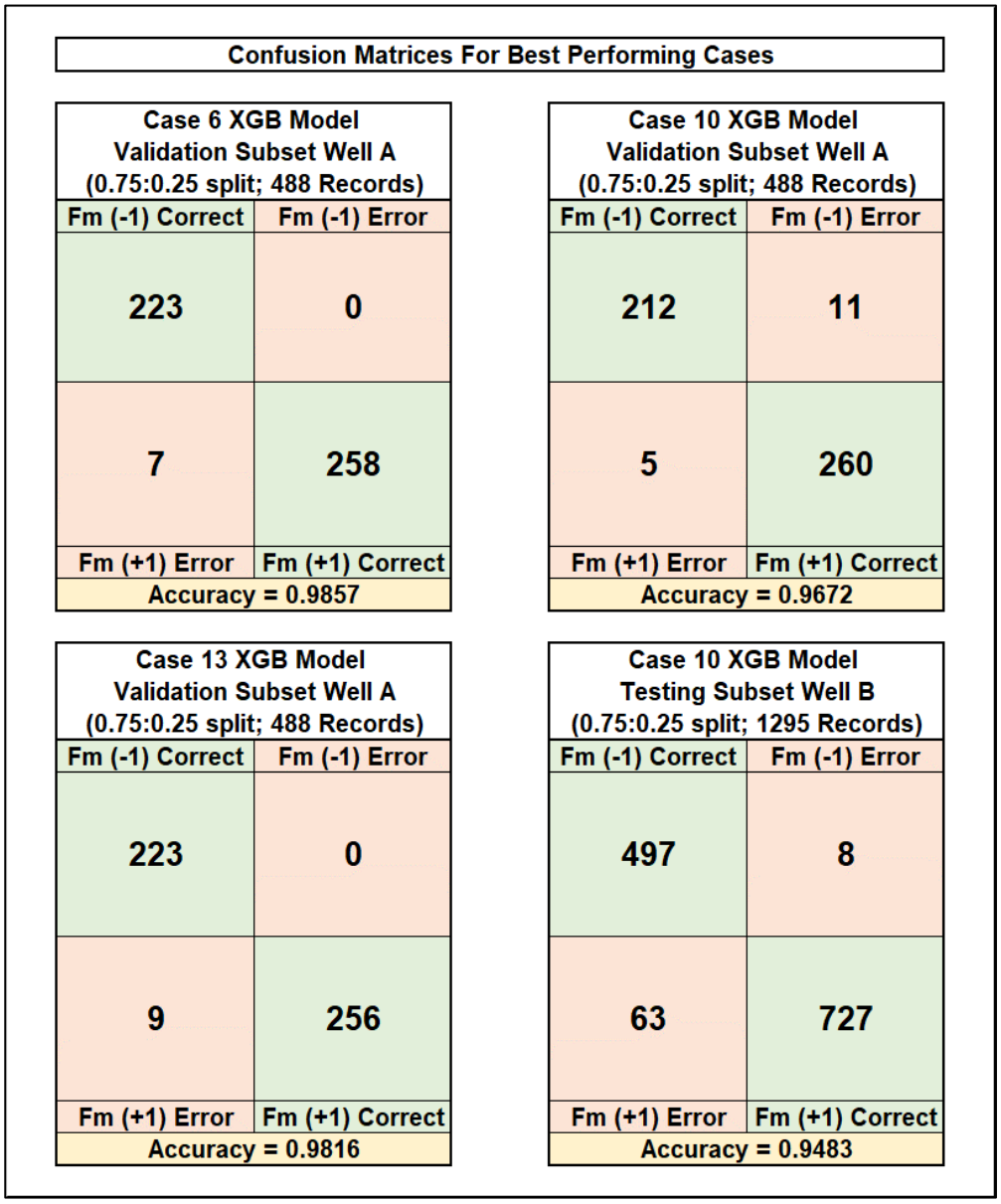

Figure S4. Confusion matrices for best performing case solutions, all of which involve optimized feature-selected well-log attributes and recorded well log combinations.

Section S6. XGB Case 10 actual versus predicted depth distributions

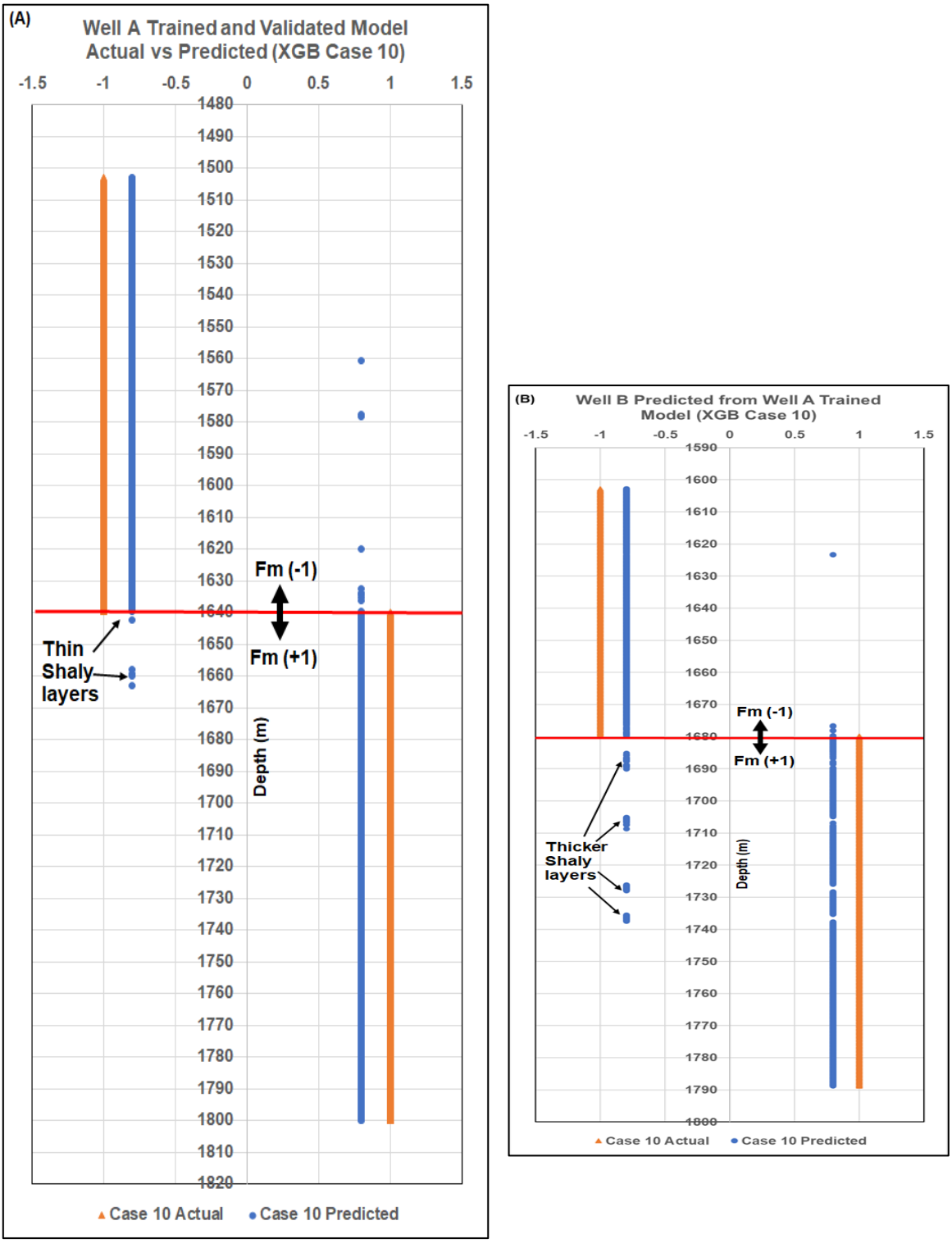

Figure S5. Case 10 XGB model results for each data record versus depth displaying actual and predicted formation categories using 0.75 : 0.25 training : testing splits. A) Well A trained /validated model; and B) Well B predictions using the Well A trained model. Note that the predicted category values are shifted by 0.2 units for display purposes to make the errors more visible.

Section S7. **Well-log attribute calculation formulas**

The method to calculate well-log derivative and volatility attributes has been described in detail elsewhere (Wood, 2022c). Six well-log attributes (G1 to G6) can be calculated generically for any recorded and quality-controlled well log (G0).

G1. First-derivative attribute

$$G1_d = (G0_d - G0_{d-1})/Abs(d - (d-1)) \quad \text{(E4)}$$

where $G0_d$ is *G*0 value at depth *d*, and $G0_{d-1}$ is *G*0 at depth *d* - 1. The depth recording interval of each well-log sample in this study is ~15 cm.

G2. Moving-average of first-derivative attribute

$$G2_{d\alpha} = (\sum_{i=1}^{i=\alpha} G1_{d-i})/\alpha \quad \text{(E5)}$$

where α is a user-defined log-sample interval covering a short depth range immediately above depth *d*. Experience suggests that α values between 3 and 10 tend to work best at picking out textural information from the recorded well log. The optimum α value varies from one formation to another depending on the frequency of log value fluctuations within it and is best determined by trial and error.

G3. Second-derivative Attribute

$$G3_{d\beta} = (G1_d - G1_{d-\beta})/Abs(d - (d-\beta)) \quad \text{(E6)}$$

where β is a user-defined log-sample interval covering a short depth range immediately depth *d*. β values is user-defined and is best determined by trial and error; values between 3 and 10 tend to perform well with a range of lithologies.

G4. Natural logarithm of ratio between adjacent log values

$$G4_{i(d)} = Ln(G0_d/G0_{d-1}) \quad \text{(E7)}$$

where $i(d)$ represents the depth interval of the calculated G4 considering both depths *d*-1 and *d*.

G5. Standard deviation of G4 for a specified overlying interval depth interval

$$G5_{i(\gamma)} = \sqrt{\frac{\sum_{j=0}^{\gamma}(G4_{i(d-j)} - G4_{i(\gamma)}mean)^2}{\gamma - 1}} \quad \text{(E8)}$$

where the attribute $G5_{i(\gamma)}$is referred to as volatility, $i(\gamma)$ represents the depth interval between recorded log depths *d*-γ and *d*. γ value is user-defined and is best determined by trial and error; values between 3 and 10 tend to perform well with a range of lithologies.

G6. Moving-average volatility

$$G6_{i(\delta)} = (\sum_{i=0}^{i=\delta} G5i_{(d-i)})/\delta \quad \text{(E9)}$$

where δ is a user-defined depth interval immediately above depth *d*. δ is assigned a value of 10 for this dataset, as determined by trial and error; values between 3 and 10 tend to perform well with a range of lithologies.

Section S8. Optimizer cost function to encourage reduced selected features

The optimizer results should ultimately be assessed and ranked in terms of their prediction accuracy. However, a key objective of the feature selection process is to seek feature combinations that involve a small number of features but generate prediction accuracy equal to or higher than that achieved by larger combination of features. This is achieved by carefully defining the fitness score (FS) or cost function used as the objective function that each optimizer attempts to minimize over a series of iterations (Wood, 2022d). The FS formula used for the KNN-optimizer analysis in this study is that defined in Eq. (E10).

$$FS = \sigma * \epsilon + \mu\,(\frac{Z}{Zmax}) \qquad \text{(E10)}$$

where σ is a user-defined constant with a value close to but just below 1, ϵ is (1- accuracy), where accuracy refers to the formation category prediction value achieved by the KNN model applying a specific feature selection expressed in terms of a fractional error, μ is 1-σ (a very small number), Z represents the number of features selected by a specific KNN solution, and Zmax represents the maximum number of features available for the KNN model to choose from. This FS configuration penalizes feature-selection solutions to a slightly greater degree the more features they select. The $\sigma * \epsilon$ accuracy-derived component tends to dominate the FS value but the $\mu\ \left(\frac{Z}{maxZ}\right)$ component adds a sufficient penalty to encourage the FS to favour solutions with fewer features selected. In the KNN-optimizer models evaluated in this study a value of σ =0.95 was applied, so μ =0.05.