International Journal of Coal Geology 249 (2022) 103903

Contents lists available at ScienceDirect

International Journal of Coal Geology

journal homepage: www.elsevier.com/locate/coal

Predicting hydrogen and oxygen indices (HI, OI) from conventional well logs using a Random Forest machine learning algorithm

John B. Gordon [a,*], Hamed Sanei [b], Per K. Pedersen [a]

[a] Department of Geoscience, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada
[b] Lithospheric Organic Carbon (LOC) Group, Department of Geoscience, Aarhus University, Høegh-Guldbergs Gade 2, building 1671, 223, 8000 Aarhus C, Denmark

ARTICLE INFO

Keywords:
Machine learning
Organic petrology
Source rock
Organic geochemistry
Random Forest

ABSTRACT

Hydrogen Index (HI) and Oxygen Index (OI) are two critical parameters for assessing the hydrocarbon potential and depositional environment of any source rocks. The most common method to measure these values is to use programmed pyrolysis on drill samples. However, this method can be very time consuming, expensive, and in many cases much of the well bore may be overlooked due to biased sampling. Geochemical parameter predictions from wireline logs (i.e., Passey) have been used in the past to varying success. This is largely because petrophysical predictions often attempt to solve for linear regression solutions where this may not be the case. Here we evaluate the use of a Random Forest (RF) machine learning (ML) model to predict HI and OI from four wells from the offshore east coast of Newfoundland, Canada. The model was trained and tested using programmed pyrolysis data, organic petrology techniques, and wireline logs for prediction. The model was evaluated using mean absolute error (MAE), root mean square error (RMSE), correlation of determination (R^2), and Spearman's rank correlation (R_2). Excellent correlation coefficients were observed for RF model predictions for HI and OI that range 0.90 to 0.98 and 0.90 to 0.95 R^2 respectively. The MAE for HI and OI values range 17.30 to 52.48 and 2.82 to 12.79 respectively. The RMSE for HI and OI range 21.43 to 71.51 and 3.85 to 16.82 respectively. The Spearman's rank correlation for HI and OI range 0.87 to 0.97 and 0.90 to 0.96 respectively. This study confirms that the use of ML models can be extremely useful to predict geochemical parameter from wireline logs.

1. Introduction

Source rock evaluation in conventional hydrocarbon systems begins by assessing the quality, quantity, and thermal maturity of organic matter (Carvajal-Ortiz and Gentzis, 2015). A typical first pass approach to assess this is by using programmed pyrolysis in an attempt to understand hydrocarbon generation and expulsion and eventual accumulation and production. A good source rock would be one that is considered to have a high total organic carbon (TOC) content, however not all organic carbon has the potential to generate hydrocarbon (Tissot et al., 1974). Organic matter must be associated with hydrogen to be able to generate significant hydrocarbon. Therefore, a source rock with high hydrogen content is desirable. Hydrogen content is estimated by programmed pyrolysis by measuring the amount of hydrocarbons formed during the thermal decomposition of organic matter in the sample. This is measured in milligrams of hydrocarbons per gram of rock and is noted in programmed pyrolysis as S2 (Espitalie et al., 1977). Hydrogen index (HI) is equal to (S2/TOC) x 100. Oxygen index (OI) which is a measure of S3 from programmed pyrolysis in milligrams of CO_2 per gram of rock is equal to (S3/TOC)x100. High OI in a source rock is indicative of gas-prone, terrigenous-sourced, kerogen type III and IV source rock and hence is considered undesirable quality for a good oil-prone source rock. The programmed pyrolysis method (e.g., Rock-Eval, HAWK, etc.) is a good first approach to evaluate hydrocarbon potential based on the aforementioned criterial, but one must use extreme caution in using programmed pyrolysis as a stand-alone dataset (e.g., Dembicki, 2009). Other datasets must be integrated in order to understand the sedimentary systems that deposited the source rock and understand the type, distribution, and preservation of the organic matter.

For many years researchers have attempted with variable success to relate geochemistry data to wireline log data in an attempt to relate these geochemical parameters to well log information to predict good source rock intervals (e.g., Passey et al., 1990, 2012; Creaney and

* Corresponding author.
E-mail address: john.gordon1@ucalgary.ca (J.B. Gordon).

https://doi.org/10.1016/j.coal.2021.103903
Received 27 July 2021; Received in revised form 8 November 2021; Accepted 14 November 2021
Available online 17 November 2021

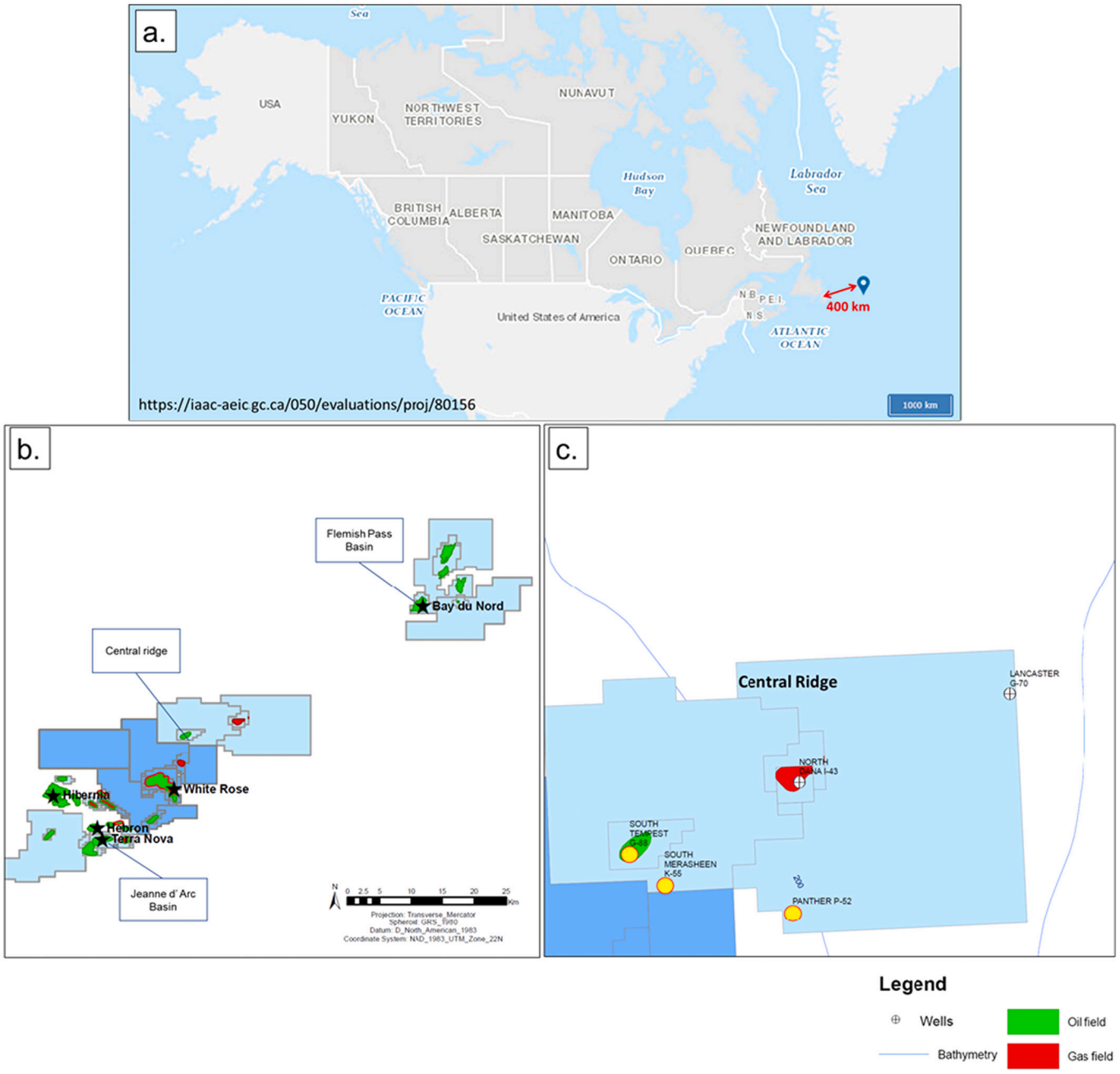


Fig. 1. . Area map. a.) Map of Canada showing the location of Newfoundland and the study area. b.) Map showing the location of the Jeanne d'Arc, Flemish Pass Basins and the Central Ridge. c.) Map showing the location of the three wells in this study from the Central Ridge (yellow dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

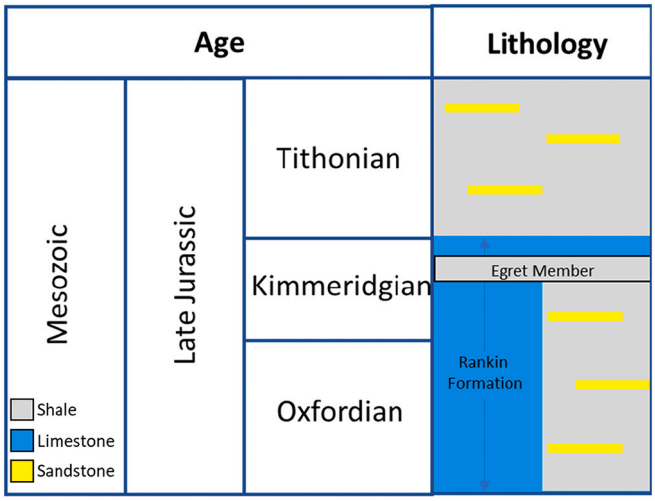


Fig. 2. Jeanne d'Arc Basin lithostratigraphic chart showing the Tithonian and Kimmeridgian intervals of interest. The Rankin Formation and the Egret Member are highlighted (Modified from Enachescu, 2005).

Passey, 1993). However, due to the complexity of comparing geochemistry data and well log data, many of the relationships are non-linear and fail to give accurate results when using only linear regression models (Wang et al., 2019a, 2019b). A relatively new approach to this problem is using machine learning algorithms than can produce more accurate results than traditional statistical analysis and can handle large datasets producing fast and cost-effective results (Rodriguez-Galiano et al., 2015).

The aim of this study is to test the predictive accuracy of Random Forest (RF) analysis, a machine learning algorithm based on an ensemble of decision trees (Breiman, 2001) to predict HI and OI using only wireline logs as input data on four east coast wells from offshore Newfoundland. Random Forest analysis has a proven track record in many scientific and engineering fields and has proven predictive accuracy in classification and regression analysis (Handhal et al., 2020). First, programmed pyrolysis, and organic petrology techniques will be used to gain an understanding of source rock types, organic matter types, and the comparison to HI and OI parameters. Second, run a machine learning Random Forest analysis using a normalized triple combo suite of well logs from the four east coast wells to predict HI and OI, Third, compare the results to the known source rock and organic matter types to test the Random Forest prediction.

2. Geological background

This study is based on three wells drilled in the Central Ridge area located off the east coast of Newfoundland approximately 400 km east of St. John's (Fig. 1). The fourth well's name and location cannot be disclosed due to the proprietary nature of the data and will be referred to simply as Well A in this study. The Central Ridge is located in the Flemish Pass sub-basin. The Flemish Pass sub-basin formed in response to Late Triassic to Paleocene rifting events that formed deep sedimentary basins bounded by faults (Creaney and Allison, 1987; Enachescu, 2005). The three Central Ridge wells were drilled conventionally between 1980 and 1988 in search of hydrocarbon bearing Jurassic and/or Cretaceous sandstone intervals with little success (Cotterill, 1987, unpublished). In the Central Ridge area, the depositional environment has been interpreted as deltaic to prodelta deposits (BeicipFranlab, 2015). The reservoir sandstone intervals are sandwiched in between thick intervals of organic-rich and/or organic-lean mudrocks (BeicipFranlab, 2015). The mudrock intervals are the focus of this study.

The Kimmeridgian-aged Egret Member of the Rankin Formation (Fig. 2) in offshore Newfoundland has been considered the primary source rock in the Jeanne d'Arc Basin and has been extensively studied in the past (Swift and Williams, 1980; Creaney and Allison, 1987; Fowler et al., 1990, 1991; Huang, 1994; Fowler and McAlpine, 1995; DeSilva, 1999; Enachescu et al., 2010, Enachescu, 2012). It has been described as an excellent hydrocarbon potential marine source rock (Swift and Williams, 1980) and is considered to be the equivalent source rock found not only in Kimmeridgian sediments, but also preserved in Tithonian-aged sediments throughout other subbasins in offshore Newfoundland including the Central Ridge with high hydrogen index, high TOC content, and low oxygen index (Enachescu, 2005; Fowler et al., 2007). The depositional environment of this source rock has been interpreted as pelagic due to the observed lack of silt and sand grains and abundance of well preserved marine algal organic matter (Raine, 2006, unpublished). Other fine grained sediments with high organic content are also present in the thick mudrock intervals but have very little hydrocarbon potential due to the abundance of allochthonous continental derived degraded and reworked organic matter suggestive of a more terrigenous derived sediment likely from a deltaic source (Raine, 2006, unpublished). These mudrocks have low HI and high OI values with variable TOC content.

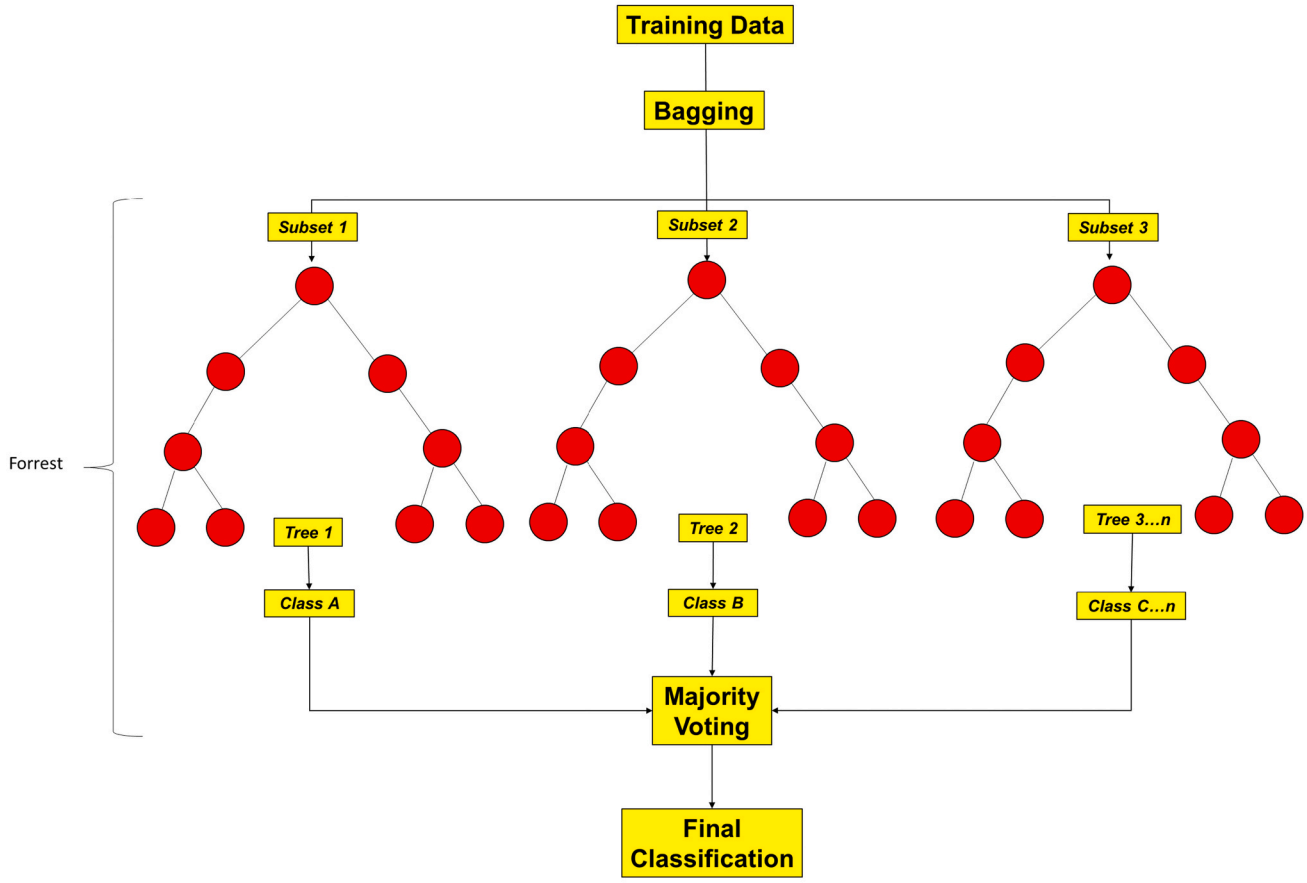
3. Sampling, data used and methodology

3.1. Programmed pyrolysis

A total of twenty-nine side wall core (SWC) samples and ninety drill cuttings samples were analyzed for hydrocarbon potential using programmed pyrolysis. The samples were sent to the Lithospheric Organic Carbon (LOC) laboratory, Department of Geoscience, Aarhus University in Denmark, for HAWK pyrolysis analysis for the standard cycle of Rock Eval 6 analysis. The samples were finely ground, and the method was carried out as described by Lafargue et al. (1998). In the analyzer, the samples are heated to an iso-temperature of 300 °C and held for 3 min followed by a ramping of the temperature by 25 °C/min to a temperature of 650 °C. S1 (free hydrocarbon) and S2 (thermally cracked kerogen) are the output parameter peaks reported in mg HC/g. The oxygen containing carbon in the kerogen released during the heating process produces the output parameter S3 and is reported as mg CO_2/g. The analyzer is then heated to a final temperature of 850 °C to oxidize the residual organic carbon and is reported as the total organic carbon (TOC: wt%) of the sample (Lafargue et al., 1998).

3.2. Organic petrology

Four SWC and nineteen drill cuttings samples were selected for organic petrology examination based on HI and OI variations from the pyrolysis data. The samples were prepared into finely polished epoxy-resin sample pellets. A Zeiss Axioimager II microscope equipped with the Diskus-Fossil system was used to carry out all of the organic petrology analysis at the GSC in Calgary, Canada. No less that 50 vitrinite reflectance measurements were taken on each sample at 50×



Software, programming environment and version

a. Data is hosted locally in the laptop.
b. Application used is Jupyter Notebook within Anaconda software.
c. Python version is 3.8
d. All codes are stored in Jupyter notebook format (*.ipynb).
e. Python libraries used are:
   i. Pandas v1.1.3 Installation: conda install pandas
   ii. Numpy v1.19.2 Installation: conda install numpy
   iii. Matplotlib v3.3.2 Installation: conda install matplotlib
   iv. Plotly v4.14.3 Installation: conda install plotly
   v. Lasio v0.28 Installation: pip install lasio
   vi. Sklearn v0.23.2 Installation: conda install scikit-learn

Fig. 3. Random Forest flow chart and software programming used in this study (Modified from Sun et al., 2019).

Table 1
List of training data for Random Forest model.

Well ID	Depth	HI	OI
	m	mg HC/gTOC	mg HC/g CO_2
Well A	3026.00	172	33
Well A	3042.10	138	36
Well A	3076.80	0	158
Well A	3077.70	81	134
Well A	3091.50	235	61
Well A	3110.00	230	81
Well A	3138.00	242	93
Well A	3178.00	186	109
Well A	3193.00	493	14
Well A	3202.10	904	2
Well A	3208.70	426	11
Well A	3212.00	515	9
Well A	3219.30	869	2
Well A	3224.00	68	94
Well A	3268.70	271	44
Well A	3289.00	57	109
Well A	3307.20	100	103
Well A	3335.00	81	142
Well A	3420.00	361	46
Well A	3496.00	55	84
Well A	3520.00	413	12
Well A	3542.00	103	155
Well A	3600.00	209	71
Well A	3612.00	110	88
Well A	3618.40	178	103
Well A	3674.10	321	17
Well A	3723.00	181	91
Well A	3769.00	221	31
Well A	3806.70	252	18
Panther P-52_E.Tith	3230.00	190	35
Panther P-52_E.Tith	3260.00	484	20
Panther P-52_E.Tith	3305.00	570	19
Panther P-52_E.Tith	3345.00	622	14
Panther P-52_E.Tith	3380.00	622	15
Panther P-52_E.Tith	3425.00	357	28
Panther P-52_E.Tith	3465.00	511	14
Panther P-52_E.Tith	3495.00	497	24
Panther P-52_E.Tith	3525.00	507	16
Panther P-52_E.Tith	3545.00	599	9
Panther P-52_E.Tith	3565.00	661	6
Panther P-52_Kimm	3755.00	180	88
Panther P-52_Kimm	3760.00	122	128
Panther P-52_Kimm	3775.00	102	128
Panther P-52_Kimm	3780.00	350	20
Panther P-52_Kimm	3790.00	147	92
Panther P-52_Kimm	3805.00	446	19
Panther P-52_Kimm	3815.00	401	14
Panther P-52_Kimm	3825.00	510	7
Panther P-52_Kimm	3830.00	455	13
Panther P-52_Kimm	3840.00	357	18
Panther P-52_Kimm	3855.00	362	19
Panther P-52_Kimm	3860.00	435	14
Panther P-52_Kimm	3870.00	293	17
Panther P-52_Kimm	3885.00	377	11
Panther P-52_Kimm	3890.00	290	21
Panther P-52_Kimm	3905.00	315	22
Panther P-52_Kimm	3915.00	252	20
Panther P-52_Kimm	3945.00	212	48
Panther P-52_Kimm	3950.00	251	32
Panther P-52_Kimm	3965.00	195	66
Panther P-52_Kimm	3975.00	188	88
Panther P-52_Kimm	3980.00	185	83
Panther P-52_Kimm	3995.00	243	36
Panther P-52_Kimm	4005.00	214	33
Panther P-52_Kimm	4014.00	187	82
Panther P-52_Kimm	4025.00	121	109
South Merasheen K-55_L. Tith.	2435.00	185	19
South Merasheen K-55_L. Tith.	2445.00	227	22
South Merasheen K-55_L. Tith.	2475.00	468	15
South Merasheen K-55_L. Tith.	2495.00	695	12
South Merasheen K-55_L. Tith.	2505.00	739	7
South Merasheen K-55_E.Tith	2545.00	366	18
South Merasheen K-55_E.Tith	2555.00	413	12

(*continued on next page*)

Table 1 (*continued*)

Well ID	Depth	HI	OI
	m	mg HC/gTOC	mg HC/g CO_2
South Merasheen K-55_E.Tith	2595.00	493	16
South Merasheen K-55_E.Tith	2635.00	479	12
South Merasheen K-55_E.Tith	2645.00	560	13
South Merasheen K-55_E.Tith	2655.00	670	11
South Merasheen K-55_E.Tith	2665.00	622	10
South Merasheen K-55_E.Tith	2675.00	675	10
South Merasheen K-55_E.Tith	2685.00	202	28
South Merasheen K-55_E.Tith	2715.00	508	19
South Merasheen K-55_E.Tith	2725.00	345	16
South Merasheen K-55_E.Tith	2745.00	147	35
South Merasheen K-55_E.Tith	2765.00	271	23
South Merasheen K-55_E.Tith	2775.00	363	24
South Merasheen K-55_E.Tith	2785.00	90	37
South Merasheen K-55_E.Tith	2795.00	249	24
South Merasheen K-55_E.Tith	2815.00	90	32
South Merasheen K-55_E.Tith	2830.00	45	65
South Merasheen K-55_Kimm	3015.00	352	22
South Merasheen K-55_Kimm	3035.00	457	22
South Merasheen K-55_Kimm	3055.00	431	20
South Merasheen K-55_Kimm	3075.00	626	9
South Merasheen K-55_Kimm	3095.00	524	19
South Merasheen K-55_Kimm	3115.00	382	13
South Merasheen K-55_Kimm	3135.00	571	10
South Merasheen K-55_Kimm	3155.00	542	8
South Merasheen K-55_Kimm	3175.00	559	12
South Merasheen K-55_Kimm	3205.00	664	12
South Merasheen K-55_Kimm	3225.00	384	19
South Merasheen K-55_Kimm	3245.00	672	7
South Merasheen K-55_Kimm	3315.00	393	20
South Merasheen K-55_Kimm	3335.00	587	11
South Merasheen K-55_Kimm	3355.00	443	13
South Merasheen K-55_Kimm	3375.00	491	12
South Merasheen K-55_Kimm	3395.00	128	35
South Merasheen K-55_Kimm	3425.00	73	32
South Tempest G-88_L.Tith	3470.00	79	49
South Tempest G-88_L.Tith	3505.00	75	60
South Tempest G-88_L.Tith	3515.00	87	54
South Tempest G-88_L.Tith	3555.00	48	44
South Tempest G-88_L.Tith	3565.00	70	54
South Tempest G-88_L.Tith	3645.00	106	42
South Tempest G-88_L.Tith	3655.00	78	54
South Tempest G-88_L.Tith	3665.00	72	91
South Tempest G-88_E.Tith	3745.00	369	14
South Tempest G-88_E.Tith	3765.00	269	15
South Tempest G-88_E.Tith	3785.00	292	20
South Tempest G-88_E.Tith	3805.00	323	13

GR	RESD	DENS	NPHI	DTC
API	ohm/m	Kg/m^3	*v*/v	μs/m
89.61	0.95	2491.17	0.48	348.01
91.80	0.93	2495.21	0.53	342.98
12.46	6.35	2725.35	0.11	210.07
29.54	3.64	2599.71	0.17	250.77
72.30	2.90	2626.30	0.32	274.80
69.37	2.09	2586.34	0.35	289.44
62.21	2.68	2619.88	0.19	241.86
42.01	3.85	2692.14	0.25	242.94
54.28	10.13	2632.38	0.25	249.51
58.61	12.22	2403.06	0.43	327.34
61.18	12.71	2574.80	0.32	285.21
42.65	16.62	2586.81	0.23	252.76
28.37	23.09	2621.18	0.15	242.23
83.04	1.51	2584.60	0.37	304.36
51.85	4.11	2544.30	0.20	253.44
92.43	1.09	2552.25	0.40	319.39
72.44	1.64	2595.02	0.35	296.43
61.15	7.45	2686.17	0.22	225.08
14.62	15.37	2506.69	0.13	220.57
77.93	2.54	2497.64	0.17	255.85
43.21	7.58	2641.89	0.12	214.32
40.40	4.35	2635.85	0.17	217.74
98.88	1.85	2547.71	0.38	274.47

(*continued on next page*)

Table 1 (*continued*)

GR	RESD	DENS	NPHI	DTC
API	ohm/m	Kg/m^3	*v*/v	μs/m
43.57	2.45	2608.12	0.26	253.38
25.62	6.49	2660.48	0.16	231.73
85.32	1.81	2535.74	0.38	317.55
78.40	2.01	2615.03	0.34	275.63
76.75	1.61	2573.72	0.38	326.27
88.88	1.55	2543.16	0.37	269.43
79.06	2.41	555.76	0.30	274.23
84.50	3.59	468.61	0.38	316.45
90.77	3.06	530.26	0.35	268.00
67.29	6.90	514.92	0.33	292.07
82.25	7.80	391.30	0.35	299.39
74.93	6.82	553.73	0.27	316.23
83.62	6.71	502.49	0.36	281.03
62.51	5.42	588.28	0.27	259.05
65.99	12.08	2518.11	0.27	264.31
70.03	13.93	2608.75	0.23	270.55
55.22	72.15	2396.61	0.27	278.06
49.07	55.70	2613.73	0.08	201.89
83.49	12.21	2674.51	0.19	225.67
86.24	12.83	2641.22	0.22	233.30
63.66	27.23	2638.17	0.16	210.36
58.23	33.38	2713.57	0.19	213.28
67.44	29.23	2682.71	0.21	235.86
61.25	37.34	2574.41	0.28	256.07
77.48	649.58	2485.82	0.29	277.58
38.07	32.71	2686.41	0.19	200.72
70.64	34.12	2626.31	0.20	267.43
67.57	268.55	2560.82	0.27	292.93
86.23	32.98	2636.24	0.24	244.88
58.60	163.17	2503.91	0.29	255.64
52.89	37.28	2700.09	0.15	185.92
88.23	50.38	2555.80	0.32	287.05
64.05	63.77	2634.32	0.23	235.51
67.92	214.30	2500.05	0.30	287.51
83.60	30.24	2570.10	0.25	268.73
80.10	47.14	2589.03	0.21	271.95
77.01	23.30	2611.43	0.24	257.80
89.67	14.00	2647.70	0.25	230.03
48.88	31.06	2638.14	0.18	214.86
50.12	58.54	2620.61	0.20	240.10
34.78	100.98	2681.83	0.10	184.05
82.63	21.73	2649.54	0.19	210.39
93.71	12.26	2671.65	0.20	218.60
69.50	1.13	2480.64	0.37	343.85
59.07	1.38	2565.75	0.31	312.19
77.40	2.01	2551.60	0.28	321.10
54.98	4.38	2562.60	0.23	263.86
73.27	1.57	2568.23	0.29	306.53
71.24	2.83	2417.03	0.38	315.73
67.74	3.41	2375.10	0.39	325.00
62.99	2.67	2554.99	0.28	269.93
68.90	2.62	2518.52	0.33	309.38
50.20	4.35	2544.84	0.25	270.70
78.34	4.17	2376.50	0.43	337.00
43.18	9.98	2329.75	0.32	307.85
86.20	2.29	2587.10	0.31	290.55
55.73	3.63	2511.69	0.24	241.82
57.88	4.09	2505.23	0.37	302.35
70.10	2.11	2586.47	0.32	266.88
85.78	1.82	2534.30	0.35	310.82
83.33	2.54	2542.30	0.28	272.41
76.36	2.46	2540.84	0.29	277.22
79.33	2.00	2597.08	0.34	281.51
82.67	1.80	2542.22	0.35	288.42
61.93	2.06	2548.78	0.18	242.58
92.57	2.12	2617.76	0.29	286.29
69.10	5.32	2722.58	0.18	238.12
78.79	6.08	2580.38	0.31	272.78
70.53	14.28	2478.50	0.27	255.76
70.69	12.09	2564.11	0.23	255.10
69.04	9.66	2749.79	0.20	229.49
67.60	9.28	2680.29	0.21	222.87
75.28	11.25	2580.35	0.28	270.61
59.13	21.58	2584.66	0.24	237.80

(*continued on next page*)

Table 1 (*continued*)

GR	RESD	DENS	NPHI	DTC
API	ohm/m	Kg/m^3	v/v	µs/m
72.52	22.42	2557.71	0.33	264.54
65.75	14.89	2508.05	0.30	281.50
77.79	6.70	2698.35	0.21	233.31
67.03	28.18	2549.06	0.31	286.40
46.61	53.97	2647.19	0.19	230.58
74.81	72.87	2424.40	0.32	350.96
81.14	7.48	2588.63	0.21	279.92
59.22	6.97	2702.25	0.13	202.35
98.48	4.90	2678.53	0.23	226.79
56.22	11.60	2719.31	0.08	191.44
33.74	1.50	2091.60	0.28	304.76
50.82	1.79	2112.40	0.42	262.96
45.90	2.60	2072.80	0.48	262.24
52.22	1.53	2958.70	0.28	297.84
47.82	2.51	2928.30	0.32	270.56
26.58	2.74	2728.20	0.51	369.24
53.87	3.12	2597.10	0.23	240.20
55.76	9.89	2622.20	0.24	243.96
65.55	3.46	2353.70	0.37	291.72
76.25	3.40	2409.80	0.35	295.00
74.96	3.68	2520.00	0.30	319.12
73.46	4.42	2557.10	0.35	289.20

Table 2
List of RF model parameters used.

bootstrap = True,	min_samples_leaf = 1,
ccp_alpha = 0.0,	min_samples_split = 2,
criterion = 'mse',	min_weight_fraction_leaf = 0.0,
max_depth = None,	n_estimators = 25,
max_features = 'auto',	n_jobs = None,
max_leaf_nodes = None,	oob_score = False,
max_samples = None,	random_state = 42, verbose = 0,
min_impurity_decrease = 0.0,	warm_start = False
min_impurity_split = None,	

magnification. An ultrafine measurement probe (0.3 µm^2 spot size) was used under oil immersion (refractive index, n = 1.518 at 23 °C). An yttrium-aluminum-garnet reference standard was used with a reflectance of 0.906% under oil immersion. Maceral point counts were carried out using a twenty-one cross-hair grid (e.g., Gordon et al., 2021). No less than two-hundred maceral counts per sample were counted to produce an organic maceral distribution normalized to 100% of the measured TOC. No macerals were counted that appeared to be isolated in the sample binder. The maceral categories (vitrinite, inertinite, liptinite, and solid bitumen) were determined based on maceral attributes described in ICCP (1998), ICCP (2001), Pickel et al., 2017, and Sanei, 2020.

3.3. Well log data

Many studies have shown that wireline logs can be sensitive to the presence of organic matter in rocks (Passey et al., 1990; Creaney and Passey, 1993; Passey et al., 2012; Bolandi et al., 2017; Wang et al., 2019a, 2019b; Handhal et al., 2020). A single log parameter may be sensitive to certain downhole conditions and lithology, therefore multiple log parameters are required for RF prediction models (Wang et al., 2019a, 2019b). The most common logs showing sensitivity to organic matter include the natural gamma, resistivity, transit interval time, and porosity (density and neutron). In general, the higher the organic content in the rock, the more obvious the response is to the wireline logs (Wang et al., 2019a, 2019b). The wireline log input data used for RF training in this study include gamma ray (GR), resistivity (RESD), density (DENS), neutron (NPHI), and sonic DTC). All wireline logs were mathematically normalized to each other using the histogram module in Interactive Petrophysics® software to improve data consistency and integrity in order to create a common basis for log comparison.

3.4. Random Forest and well log data

3.4.1. Random Forest modelling

Random Forest (RF) is a machine learning method that consists of an ensemble of randomized classification and regression trees (CART) that generates many decision trees to improve the performance of the prediction model (Breiman, 2001). A decision tree represents a set of limits that are hierarchically organized and randomly applied from a root node as many times as the number of trees in the ensemble (Keykhay-Hosseinpoor et al., 2020). The trees are generated from a subset of training samples through replacement (a bagging approach) and the same sample can be selected several times or may not be selected at all. Approximately two thirds of the samples (referred to as in-bag samples) are used to train the trees and the remaining one third (referred to as out-of-the bag samples) are used to confirm how well the RF model performs (Breiman, 2001). The final classification decision, or a committee vote, is taken by averaging (using the arithmetic mean) the class assignment probabilities calculated by all produced trees (Belgiu and Drăguţ, 2016). The main advantage of using this approach is the RF classification algorithm can model non-linear relationships and the model consists of numerous random decisions trees. Each individual tree creates an uncorrelated forest of trees whose prediction by voting committee is more accurate (Grimm et al., 2008). The RF workflow and software used in

Table 3
Programmed pyrolysis results.

Well ID	Depth	S1	S2-Kerogen	S3	Tmax-Maturity	TOC-Total Organic Carbon
	m	mgHC/g	mgHC/g	$mgCO_2/g$	°C	wt%
Well A	3026.0	0.16	2.82	0.54	428	1.64
Well A	3042.1	0.10	2.13	0.56	429	1.54
Well A	3064.0	0.13	19.43	0.43	429	3.76
Well A	3076.8	0.00	0.00	0.44	438	0.28
Well A	3077.7	0.03	0.29	0.48	439	0.36
Well A	3091.5	0.11	2.43	0.64	426	1.03
Well A	3110.0	0.07	2.20	0.77	429	0.95
Well A	3138.0	0.08	1.91	0.74	431	0.79
Well A	3178.0	0.11	1.15	0.68	433	0.62
Well A	3193.0	0.16	8.40	0.24	426	1.70
Well A	3202.1	0.23	135.26	0.45	420	14.96
Well A	3208.7	0.16	6.19	0.17	424	1.45
Well A	3212.0	0.96	10.94	0.21	427	2.12
Well A	3219.3	0.17	71.58	0.22	426	8.23
Well A	3224.0	0.28	0.59	0.81	432	0.86
Well A	3268.7	1.20	3.65	0.60	426	1.34
Well A	3289.0	0.13	0.58	1.10	430	1.01
Well A	3307.2	0.16	0.93	0.97	434	0.93
Well A	3335.0	0.20	0.35	0.62	426	0.43
Well A	3394.0	0.09	0.71	0.93	433	1.07
Well A	3420.0	0.06	2.50	0.32	434	0.69
Well A	3483.2	0.09	1.36	0.71	433	1.06
Well A	3496.0	0.17	0.32	0.48	428	0.57
Well A	3520.0	0.12	7.65	0.24	430	1.85
Well A	3542.0	0.06	0.57	0.84	433	0.54
Well A	3600.0	0.08	2.04	0.69	432	0.97
Well A	3612.0	0.05	0.68	0.55	430	0.62
Well A	3618.4	0.05	0.94	0.54	432	0.52
Well A	3674.1	0.21	7.48	0.40	438	2.33
Well A	3723.0	0.15	1.98	1.00	431	1.09
Well A	3731.0	0.02	0.30	0.43	432	0.37
Well A	3733.7	0.02	0.08	0.53	433	0.41
Well A	3752.8	0.56	9.27	0.53	441	2.18
Well A	3769.0	0.15	3.37	0.48	434	1.52
Well A	3806.7	0.08	5.81	0.42	435	2.30
Panther P-52_E.Tith	3230.0	0.29	2.32	0.43	438	1.22
Panther P-52_E.Tith	3260.0	0.55	7.47	0.32	432	1.54
Panther P-52_E.Tith	3305.0	1.46	16.32	0.56	427	2.86
Panther P-52_E.Tith	3345.0	2.16	24.97	0.59	428	4.01
Panther P-52_E.Tith	3380.0	1.71	23.72	0.57	435	3.81
Panther P-52_E.Tith	3425.0	0.73	6.10	0.49	440	1.71
Panther P-52_E.Tith	3465.0	1.95	16.82	0.48	434	3.29
Panther P-52_E.Tith	3495.0	1.37	13.62	0.68	430	2.74
Panther P-52_E.Tith	3525.0	1.58	15.37	0.51	433	3.03
Panther P-52_E.Tith	3545.0	2.86	25.65	0.42	432	4.28
Panther P-52_E.Tith	3565.0	4.42	54.15	0.54	434	8.18
Panther P-52_Kimm	3755.0	0.71	2.37	1.17	435	1.32
Panther P-52_Kimm	3760.0	0.59	1.59	1.67	434	1.30
Panther P-52_Kimm	3775.0	0.56	1.13	1.41	436	1.10
Panther P-52_Kimm	3780.0	1.42	13.86	0.79	424	3.95
Panther P-52_Kimm	3790.0	0.67	2.17	1.37	437	1.47
Panther P-52_Kimm	3805.0	2.21	20.90	0.89	432	4.68
Panther P-52_Kimm	3815.0	2.17	19.93	0.70	436	4.97
Panther P-52_Kimm	3825.0	4.30	44.00	0.67	431	8.62
Panther P-52_Kimm	3830.0	2.02	18.86	0.54	437	4.14
Panther P-52_Kimm	3840.0	2.46	13.87	0.71	438	3.88
Panther P-52_Kimm	3855.0	1.79	12.01	0.64	438	3.31
Panther P-52_Kimm	3860.0	3.16	24.47	0.79	437	5.61
Panther P-52_Kimm	3870.0	2.01	9.54	0.56	438	3.25
Panther P-52_Kimm	3885.0	3.38	19.16	0.60	438	5.08
Panther P-52_Kimm	3890.0	2.23	9.66	0.70	439	3.32
Panther P-52_Kimm	3905.0	2.45	10.59	0.75	438	3.36
Panther P-52_Kimm	3915.0	3.86	11.89	0.97	441	4.71
Panther P-52_Kimm	3945.0	1.50	3.89	0.89	441	1.83
Panther P-52_Kimm	3950.0	1.90	6.01	0.78	438	2.39
Panther P-52_Kimm	3965.0	1.26	3.62	1.23	440	1.85
Panther P-52_Kimm	3975.0	1.24	3.08	1.45	440	1.64
Panther P-52_Kimm	3980.0	1.25	3.10	1.40	440	1.67
Panther P-52_Kimm	3995.0	1.69	5.09	0.77	440	2.10
Panther P-52_Kimm	4005.0	1.66	5.94	0.92	440	2.77
Panther P-52_Kimm	4014.0	1.03	2.45	1.08	439	1.31
Panther P-52_Kimm	4025.0	0.56	1.17	1.06	441	0.97
South Merasheen K-55_L. Tith.	2435.0	0.07	3.20	0.34	434	1.73

(*continued on next page*)

Table 3 (*continued*)

Well ID	Depth	S1	S2-Kerogen	S3	Tmax-Maturity	TOC-Total Organic Carbon	HI-hydrogen index	OI-oxygen index
	m	mgHC/g	mgHC/g	$mgCO_2/g$	°C	wt%	mgHC/gTOC	$mgCO_2/gTOC$
South Merasheen K-55_L. Tith.	2445.0	0.08	4.04	0.41	431	1.78	172	33
South Merasheen K-55_L. Tith.	2475.0	0.11	11.74	0.39	427	2.50	138	36
South Merasheen K-55_L. Tith.	2495.0	0.17	30.00	0.52	422	4.31	516	11
South Merasheen K-55_L. Tith.	2505.0	0.26	48.87	0.51	420	6.61	0	158
South Merasheen K-55_E.Tith	2545.0	0.08	7.57	0.37	428	2.06	81	134
South Merasheen K-55_E.Tith	2555.0	0.10	13.40	0.39	428	3.24	235	61
South Merasheen K-55_E.Tith	2595.0	0.10	14.53	0.48	425	2.95	230	81
South Merasheen K-55_E.Tith	2635.0	0.11	14.49	0.38	425	3.02	242	93
South Merasheen K-55_E.Tith	2645.0	0.14	19.22	0.48	419	3.43	186	109
South Merasheen K-55_E.Tith	2655.0	0.16	29.08	0.50	418	4.34	493	14
South Merasheen K-55_E.Tith	2665.0	0.18	24.88	0.40	418	3.99	904	2
South Merasheen K-55_E.Tith	2675.0	0.21	30.13	0.46	421	4.46	426	11
South Merasheen K-55_E.Tith	2685.0	0.07	2.91	0.41	427	1.44	515	9
South Merasheen K-55_E.Tith	2715.0	0.15	15.70	0.60	426	3.09	869	2
South Merasheen K-55_E.Tith	2725.0	0.09	6.40	0.30	426	1.85	68	94
South Merasheen K-55_E.Tith	2745.0	0.09	1.88	0.45	430	1.27	271	44
South Merasheen K-55_E.Tith	2765.0	0.11	4.35	0.38	427	1.61	57	109
South Merasheen K-55_E.Tith	2775.0	0.10	5.63	0.38	427	1.55		
South Merasheen K-55_E.Tith	2785.0	0.07	0.83	0.34	433	0.91		
South Merasheen K-55_E.Tith	2795.0	0.13	6.10	0.61	430	2.45		
South Merasheen K-55_E.Tith	2815.0	0.09	1.39	0.50	435	1.53		
South Merasheen K-55_E.Tith	2830.0	0.05	0.27	0.39	431	0.60		
South Merasheen K-55_Kimm	3015.0	0.09	4.44	0.28	426	1.26		
South Merasheen K-55_Kimm	3035.0	0.09	6.30	0.31	426	1.38		
South Merasheen K-55_Kimm	3055.0	0.08	6.26	0.30	427	1.45		
South Merasheen K-55_Kimm	3075.0	0.13	20.68	0.32	427	3.30		
South Merasheen K-55_Kimm	3095.0	0.10	11.33	0.41	427	2.16		
South Merasheen K-55_Kimm	3115.0	0.10	7.98	0.28	426	2.09		
South Merasheen K-55_Kimm	3135.0	0.15	20.48	0.36	426	3.58		
South Merasheen K-55_Kimm	3155.0	0.11	11.93	0.18	427	2.20		
South Merasheen K-55_Kimm	3175.0	0.10	14.56	0.32	428	2.60		
South Merasheen K-55_Kimm	3205.0	0.14	23.39	0.44	426	3.52		
South Merasheen K-55_Kimm	3225.0	0.10	6.31	0.31	429	1.64		
South Merasheen K-55_Kimm	3245.0	0.14	28.13	0.31	427	4.18		
South Merasheen K-55_Kimm	3315.0	0.10	6.83	0.35	431	1.73		
South Merasheen K-55_Kimm	3335.0	0.14	17.37	0.34	433	2.96		
South Merasheen K-55_Kimm	3355.0	0.11	8.57	0.26	443	1.93		
South Merasheen K-55_Kimm	3375.0	0.12	10.80	0.28	442	2.20		
South Merasheen K-55_Kimm	3395.0	0.07	1.15	0.32	445	0.89		
South Merasheen K-55_Kimm	3425.0	0.08	0.76	0.34	442	1.04		
South Tempest G-88_L.Tith	3470.0	0.15	0.65	0.41	438	0.82		
South Tempest G-88_L.Tith	3505.0	0.18	0.75	0.60	441	0.99		
South Tempest G-88_L.Tith	3515.0	0.17	0.83	0.52	438	0.95		
South Tempest G-88_L.Tith	3555.0	0.11	0.52	0.47	441	1.06		
South Tempest G-88_L.Tith	3565.0	0.18	0.70	0.54	442	0.99		
South Tempest G-88_L.Tith	3645.0	0.35	1.23	0.49	439	1.16		
South Tempest G-88_L.Tith	3655.0	0.37	0.90	0.63	444	1.15		
South Tempest G-88_L.Tith	3665.0	0.20	0.63	0.79	443	0.87		
South Tempest G-88_E.Tith	3745.0	3.72	13.57	0.52	435	3.68		
South Tempest G-88_E.Tith	3765.0	2.04	5.83	0.34	444	2.16		
South Tempest G-88_E.Tith	3785.0	1.80	6.36	0.46	444	2.17		
South Tempest G-88_E.Tith	3805.0	3.11	9.40	0.39	443	2.91		

(*continued on next page*)

Table 3 (*continued*)

HI-hydrogen index	OI-oxygen index
mgHC/gTOC	$mgCO_2$/gTOC
100	103
81	142
66	86
361	46
128	66
55	84
413	12
103	155
209	71
110	88
178	103
321	17
181	91
82	114
20	128
425	24
221	31
252	18
190	35
484	20
570	19
622	14
622	15
357	28
511	14
497	24
507	16
599	9
661	6
180	88
122	128
102	128
350	20
147	92
446	19
401	14
510	7
455	13
357	18
362	19
435	14
293	17
377	11
290	21
315	22
252	20
212	48
251	32
195	66
188	88
185	83
243	36
214	33
187	82
121	109
185	19
227	22
468	15
695	12
739	7
366	18
413	12
493	16
479	12
560	13
670	11
622	10
675	10
202	28
508	19
345	16
147	35
271	23
363	24

(*continued on next page*)

Table 3 (*continued*)

HI-hydrogen index	OI-oxygen index
mgHC/gTOC	$mgCO_2$/gTOC
90	37
249	24
90	32
45	65
352	22
457	22
431	20
626	9
524	19
382	13
571	10
542	8
559	12
664	12
384	19
672	7
393	20
587	11
443	13
491	12
128	35
73	32
79	49
75	60
87	54
48	44
70	54
106	42
78	54
72	91
369	14
269	15
292	20
323	13

this study is shown in Fig. 3. Table 1 shows all the training data from programmed pyrolysis and wireline log data used as input data.

3.4.2. RF feature selection and model evaluation

In this study a randomly selected twenty-five trees were used, and the input matrix (training data) are the well log curves used for each well (gamma ray, resistivity, density, neutron, and sonic). The default setting for the RF prediction model is shown in Table 2. In previous studies the software parameters used can be set to adjustable user settings, however it is recommended to simply use software default settings as a first approach. Changing the model settings to higher randomness have indicated noise variables in final prediction data (e.g., Svetnik et al., 2003; Grimm et al., 2008). However, it should be noted that hyper-parameter tuning can be applied in an attempt to improve the RF performance. Probst et al., 2019 have provided the tuneRanger R package that aids in tuning the RF model with model-based optimization (MBO). To evaluate the RF model, four statistical metrics were used: i) mean absolute error (MAE), ii) root mean squared error (RMSE), iii) the correlation of determination coefficient (R^2), and iv) a Spearman's rank correlation (R_s) was used to determine the degree in which the data sets are correlated. A correction factor (c_f) was applied to Σd^2 when ranks were found to be tied. The correction factor was added to Σd^2 for each tied rank in the datasets. The MAE and RMSE are used to recognize the outliers in the dataset. R^2 is used to evaluate the accuracy of the model where x_i and y_i are the measured and estimated values of HI and OI, $\bar{x}$ and $\bar{y}$ are their arithmetic mean, and n is the total number of measured HI and OI data points.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{1}$$

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

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2 \tag{3}$$

$$R_s = 1 - \left(\frac{6\Sigma d^2}{n^3 - n} \right) \tag{4}$$

$$c_f = \frac{m(m^2 - 1)}{12} \tag{5}$$

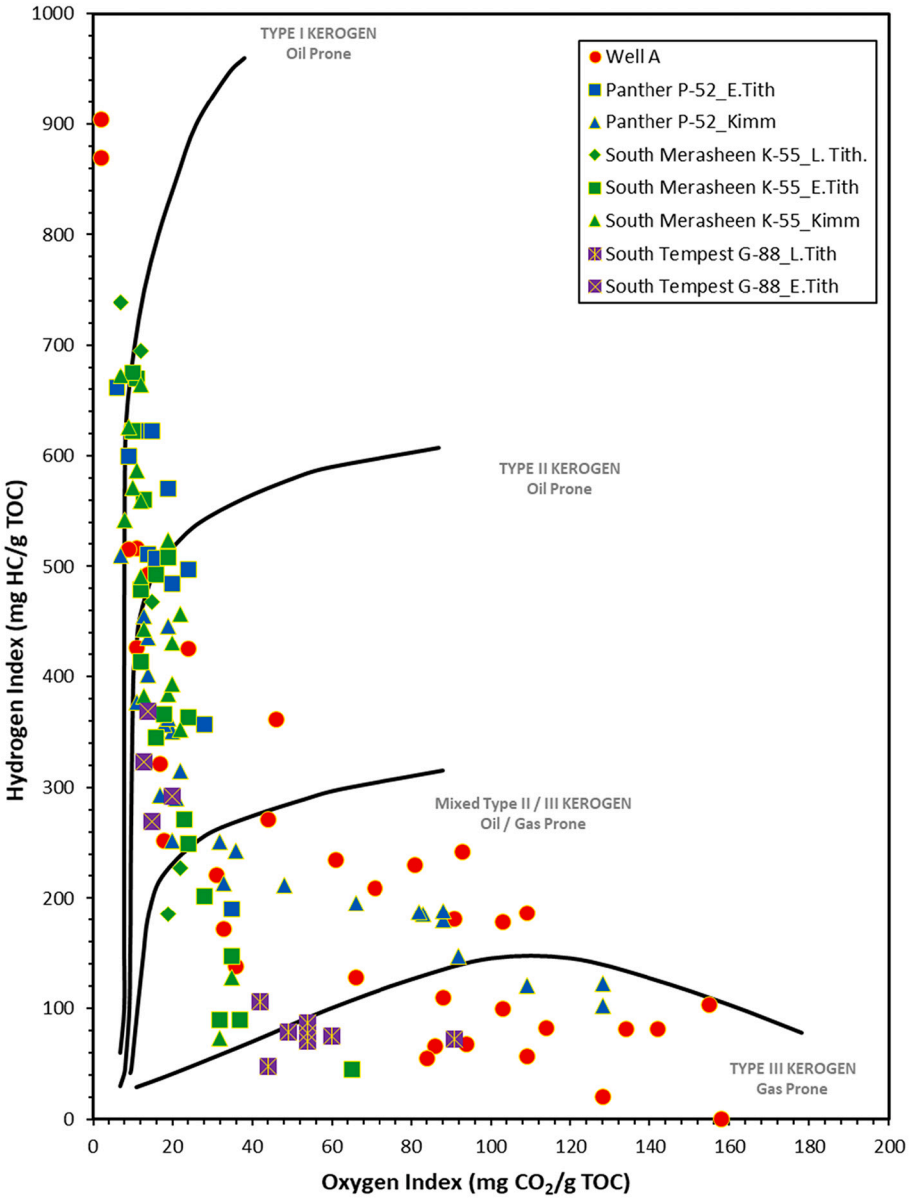


Fig. 4. Pseudo van Krevelen plot showing hydrogen index (HI) vs. oxygen index (OI) for all samples from the studied area.

Table 4
Organic petrology data collected on selected samples.

Sample information		
Well ID	Depth m	Drilling fluid
Well A	3202.1	N/A
Well A	3212.0	N/A
Well A	3289.0	N/A
Well A	3806.7	N/A
South Tempest G-88	3515.0	Water Based Mud
South Tempest G-88	3645.0	Water Based Mud
South Tempest G-88	3745.0	Water Based Mud
South Tempest G-88	3805.0	Water Based Mud
Panther P-52	3230.0	Water Based Mud
Panther P-52	3260.0	Water Based Mud
Panther P-52	3305.0	Water Based Mud
Panther P-52	3425.0	Water Based Mud
Panther P-52	3465.0	Water Based Mud
Panther P-52	3565.0	Water Based Mud
Panther P-52	3825.0	Water Based Mud
Panther P-52	3905.0	Water Based Mud
Panther P-52	4025.0	Water Based Mud
South Merasheen K-55	2445.0	Oil Based Mud
South Merasheen K-55	2505.0	Oil Based Mud
South Merasheen K-55	2545.0	Oil Based Mud
South Merasheen K-55	2675.0	Oil Based Mud
South Merasheen K-55	3225.0	Oil Based Mud
South Merasheen K-55	3245.0	Oil Based Mud

Maceral point count data (normalized to TOC)							
Vitrinite %	Reworked vitrinite %	Inertinite %	Liptinite %	Degraded liptinite %	Bitumenite %	Solid Bitumen %	Total %
0	1	1	95	0	2	0	100
8	1	7	68	0	15	2	100
52	8	1	24	0	9	6	100
23	0	9	60	0	9	1	100
1	19		76	4			100
2	31		59	8			100
1	10		83	6			100
5	33	2	19	38	3	3	100
5	45		40	9	1	1	100
1	10		86	3			100
7	10	10	60	12	1	1	100
8	26	1	61	4			100
4	7		88	1			100
2	8		84	6			100
	3		94	3			100
4	16	3	33	44			100
8	57	1	20	11	3	3	100
2	34		35	29			100
5	13		61	21			100
3	49		37	11			100
2	5	1	92	1			100
3	32		41	23	1	1	100
4	1		90	4	1	1	100

Vitrinite reflectance	
Calculated %Ro (0.0180 x Tmax) - 7.16	Measured %Ro
0.40	0.25
0.53	0.60
0.58	0.87
0.67	0.54
0.72	0.76
0.74	0.77
0.67	0.79
0.81	0.79
0.72	0.72
0.62	0.62
0.53	0.72
0.76	0.70
0.65	0.66
0.65	0.67
0.60	0.65
0.72	0.64
0.78	0.79

(*continued on next page*)

Table 4 (*continued*)

Vitrinite reflectance	
Calculated %Ro (0.0180 x Tmax) - 7.16	Measured %Ro
0.60	0.74
0.40	0.65
0.54	0.82
0.42	0.70
0.56	0.74
0.53	0.64

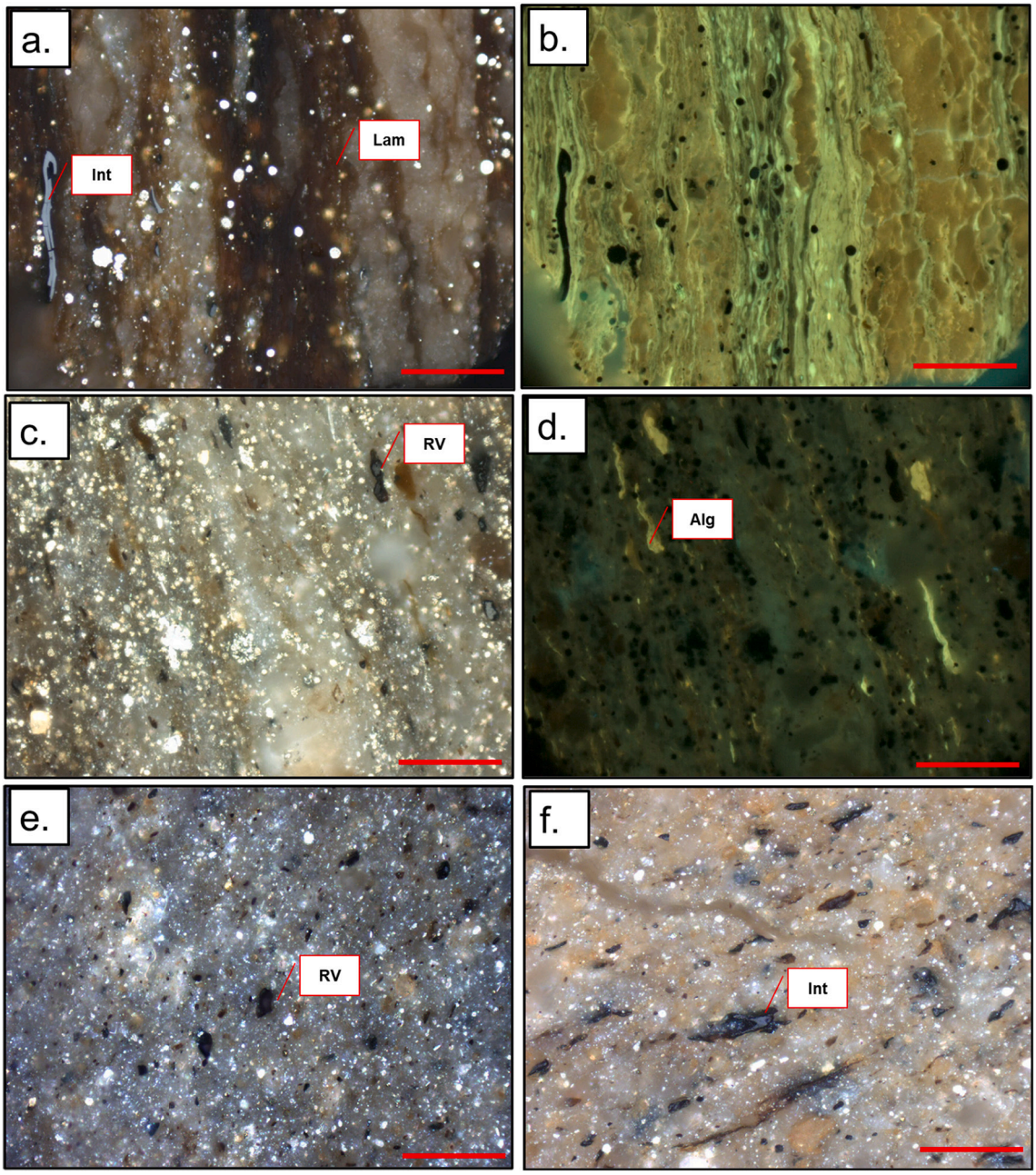


Fig. 5. Photomicrographs showing the different organic matter types. All of the photomicrographs were taken under incident-light with oil immersion and a 50 X objective. Photomicrographs a., c. e. and f. are taken under white light; photomicrographs b., and d. are taken under UV-fluorescence mode. The red scale bar represents 50 um in length. a.) Dark brown thick layers of lamalginite (Lam) from Well A. Large fragment of inertinite is also present (Int) b.) Same field of view as in a. but under fluorescence light showing bright green fluorescing algae. c.) Example from Panther P-52 showing only scattered algae fragments (Alg) with recycled vitrinite (RV) fragments and inertinite (Int) in a silty argillaceous matrix. d.) As in photo c. but under fluorescence light. e.) An example from South Tempest G-88 showing scattered recycled vitrinite (RV) and inertinite fragments (Int) is a silty argillaceous matrix. f.) Similar example from South Tempest G-88 showing recycled vitrinite (RV) and inertinite (Int) fragments. No marine algae are present in these samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

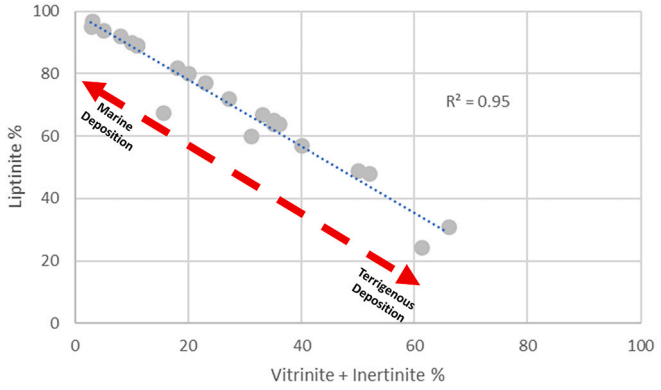


Fig. 6. Plot showing the percentage volume distribution of liptinites versus vitrinite + inertinite, as measured by microscopic point counting. Increase in distribution of liptinites suggest change to the marine depositional environment.

3.4.3. Data loading

The rock data is compiled and hosted on a local computer drive in a Microsoft Excel file format and the well log data are stored as LAS files. The application software used for RF is Jupyter Notebook within Anaconda (Python version 3.8). All of the RF code are stored in Jupyter notebook as .jpynb format. The Python libraries used are shown in Fig. 3. The Excel and LAS data is loaded with the panda read_excel function and the with the lasio_read function into separate data frames. A class object is constructed for each well to hold information such as well name, LAS filename, and curves needed for training and stored in a dictionary. The model is evaluated in scikit_learn. All of the pyrolysis data below 400 °C Tmax were filtered out.

4. Results and discussion

4.1. Programed pyrolysis and organic petrology

All data collected from programmed pyrolysis is presented in Table 3. Total organic carbon (TOC) content from all the samples collected ($n = 119$) range from 0.28 to 14.96 (wt%). S2 vales range from 0.00 to 135.26 (mg HC/g). S1 values were not used in this study as various drilling fluids were utilized during drilling that can add contamination errors in measured S1 values. The HI and OI values range from 0 to 904 (mg HC/g TOC) and from 2 to 158 (mg CO_2 /g TOC), respectively. Tmax values range from 418 to 445 °C.

The pseudo van Krevelen plot (Fig. 4) for all the samples shows a wide range of OM types from Type I oil prone, Type II oil prone, Type II/III mixed, and Type III gas prone. Pseudo van Krevelen diagrams typically show a trend for organic matter that shows a depletion of HI and OI along the defined kerogen type lines, irrespective of the organic matter type, due to thermal maturity of the organic matter (Peters et al., 2015). The samples from the four wells in this study show a progressive depletion of HI combined with increase in OI that is in contrast to the typical trend. The OM preservation here is controlled by oscillation in sea level from a marine carbonate-rich influenced OM maceral composition to more siliciclastic-rich terrigenous littoral and/or deltaic influenced OM composition. This OM composition oscillates from a liptinitic-rich and carbonate-rich aquatic depositional environment with abundant production and exceptional preservation of hydrogen-rich algal matter to a more oxygen-rich OM aquatic depositional environment typical of a terrigenous sediment input with recycled and reworked OM along with silt and sand-sized siliciclastic minerals (Omura and Hoyanagi, 2004; Hackley et al., 2020; Gordon et al., 2021).

Organic petrology results show a wide variety of maceral types are present in all the wells and age intervals including vitrinite (Type III), high-reflecting reworked vitrinite (Type IV) exhibiting a brighter grey colour than vitrinite, liptinite (Type II), inertinite (Type IV) exhibiting bright grey color and high %VRo, and solid bitumen (Table 4). These data are consistent with the programmed pyrolysis results. Representative photomicrographs of the organic petrology are illustrated in Fig. 5a to f.

Maceral point count data are shown in Table 4 and the data are normalized to measured TOC values. By plotting the sum of the relative volume of primary and reworked vitrinite plus inertinite (V + I) versus the relative volume of liptinite macerals (L) two distinct end members are showing the terrigenous influenced depositional environment versus marine. These have a strong correlation coefficient ($R^2 = 0.95$) and the strong influence of depositional environment and the oscillations in relative sea level have on the composition, preservation, and distribution of OM (Fig. 6). These observations occur regardless of the depth or age of the sediment and appear in a cyclical pattern throughout the Tithonian and Kimmeridgian.

4.2. Random Forest analysis for source rock prediction

4.2.1. Data integration

To better aid the Random Forest algorithm to predict the best source rock intervals using the well logs based on measured HI and OI data, the samples described above can be divided into three simplified groups based on the pyrolysis data and organic petrology results. TOC, HI, and OI values define these groups as they either have i) excellent hydrocarbon potential, ii) transition, and iii) poor hydrocarbon potential (Fig. 7).

4.2.2. Excellent hydrocarbon potential ($n = 13$)

Pyrolysis data for this sample grouping shows HI ranging 622 to 904 (mg HC/g TOC), OI ranging 2.0 to 15.0 (mg CO_2/g TOC), and TOC ranging 3.30 to 14.96 (wt%; Fig. 7). These samples exhibit fine laminations consisting of layers of calcite and OM. Fine disseminated pyrite crystals are associated with OM. Silt and sand-sized siliciclastic grains are rare and minor amounts of calcispheres and forams are also present indicating fully marine depositional environment. Point count data ($n = 5$) shows liptinite to be the main OM present (92.0 to 95.0%). Minor amounts of reworked vitrinite and inertinite (3.0 to 18.0%), and bituminite (0.0 to 2.0%) are also present. Fine layers of bright yellow-green algae composed of thin-walled colonial or unicellular algae occur as distinct laminae. The fluorescence color of the lamalginite in these samples is consistent with early oil window maturity. Representative photomicrographs of the point counted maceral types are illustrated in Fig. 5a and b.

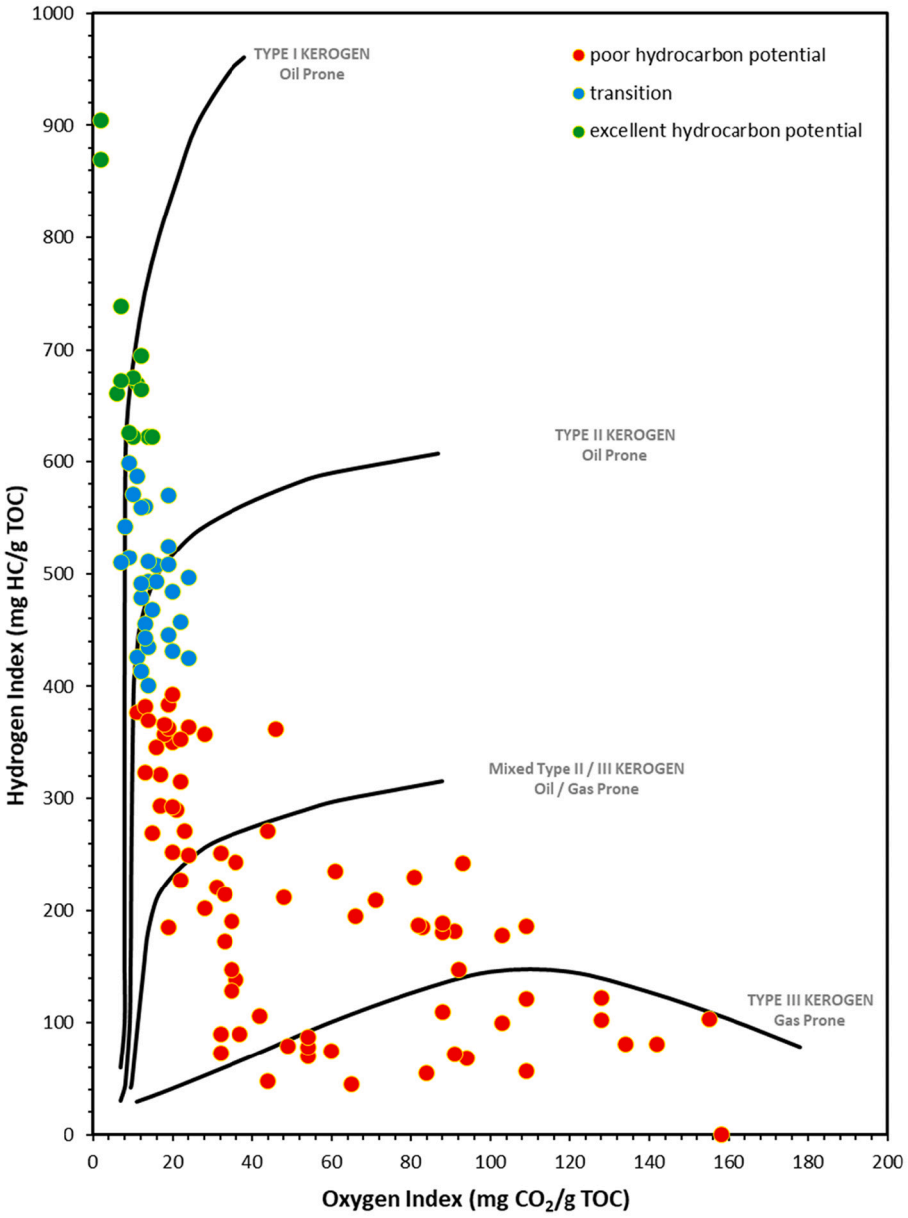


Fig. 7. Hydrogen index (HI) vs. oxygen index (OI) plot showing three groups defined by programmed pyrolysis: i) excellent hydrocarbon potential, ii) transition, and iii) poor hydrocarbon potential.

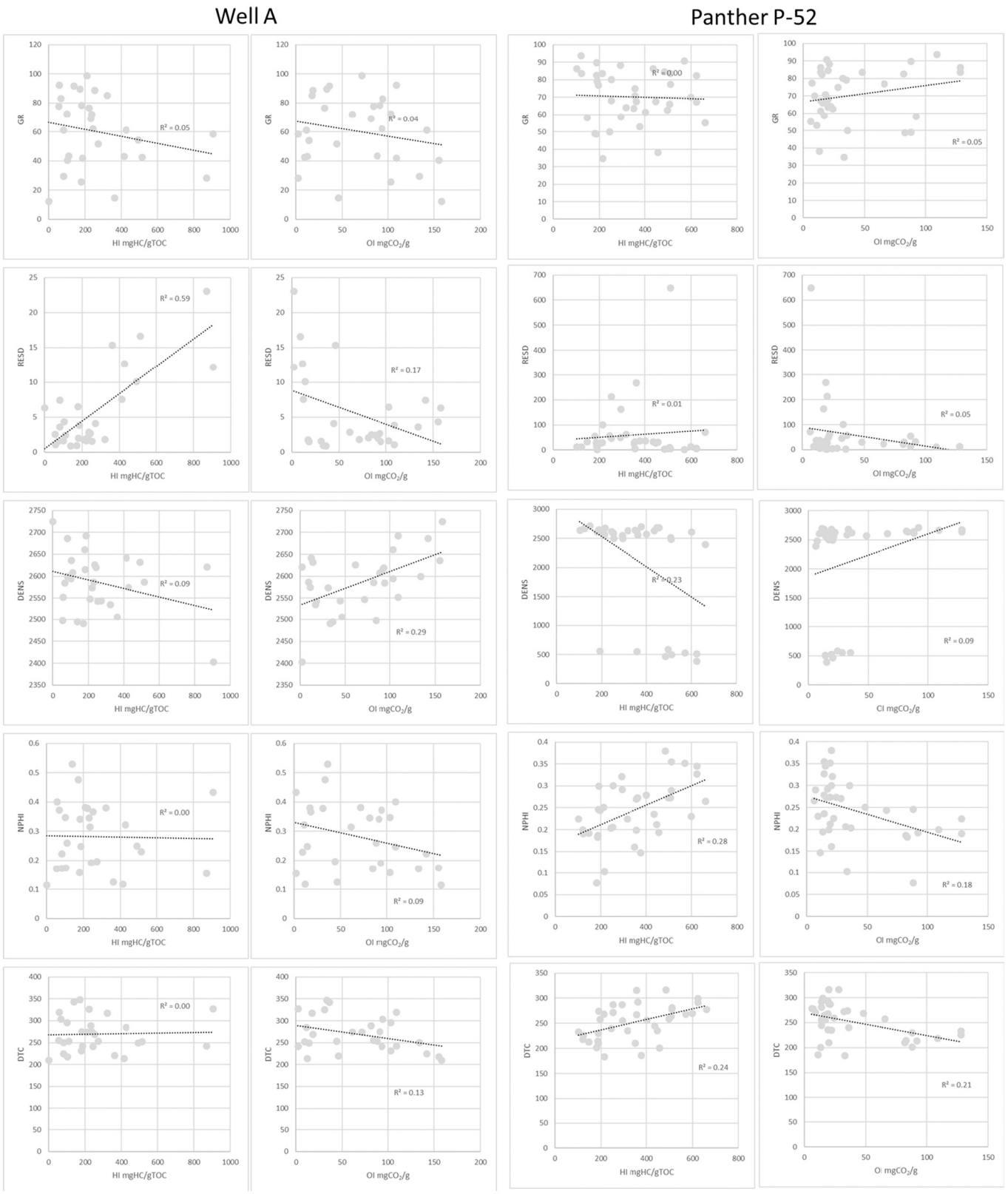


Fig. 8. Matrix scatter plots of HI and OI with various wireline logs.

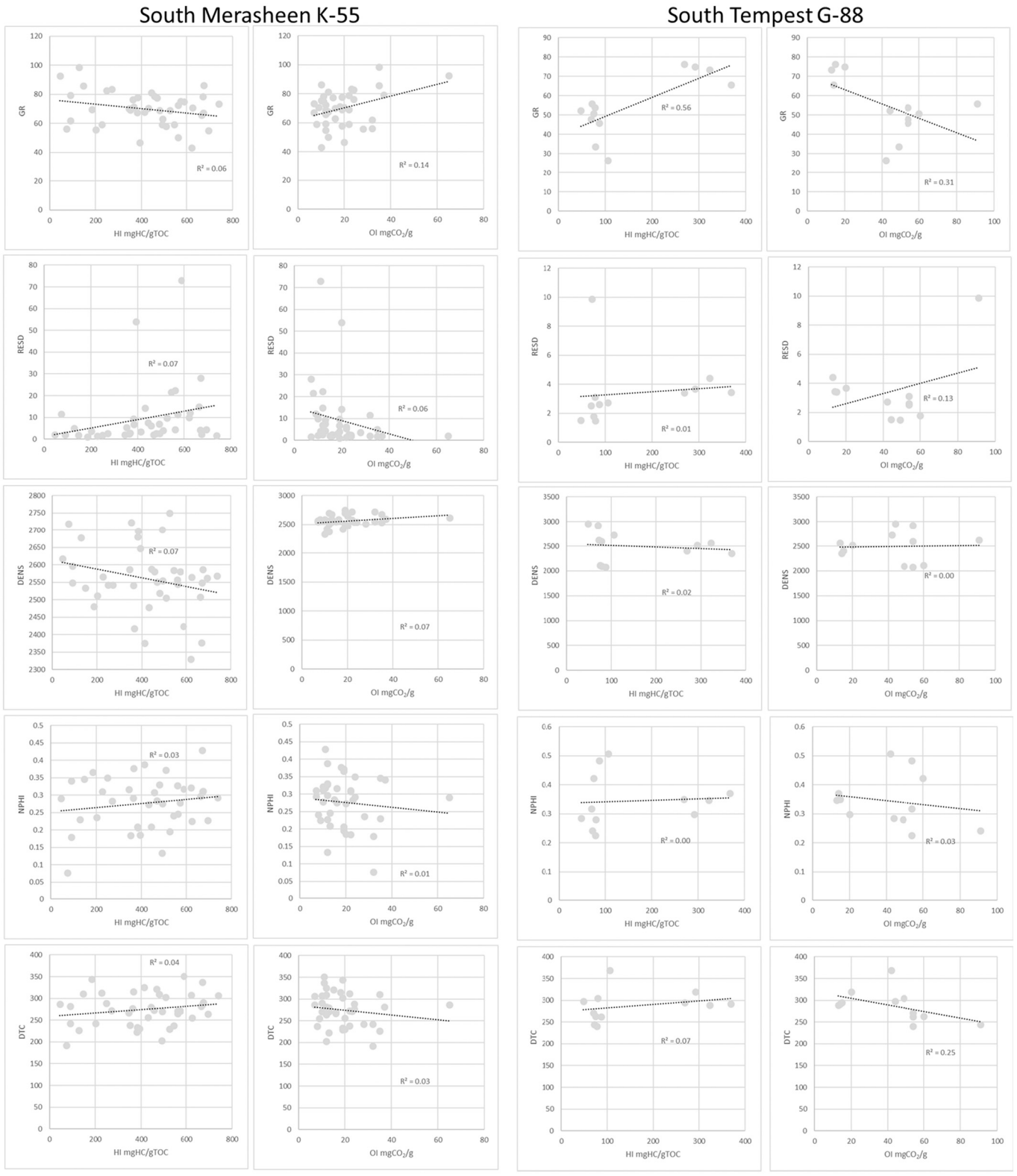


Fig. 8. (*continued*).

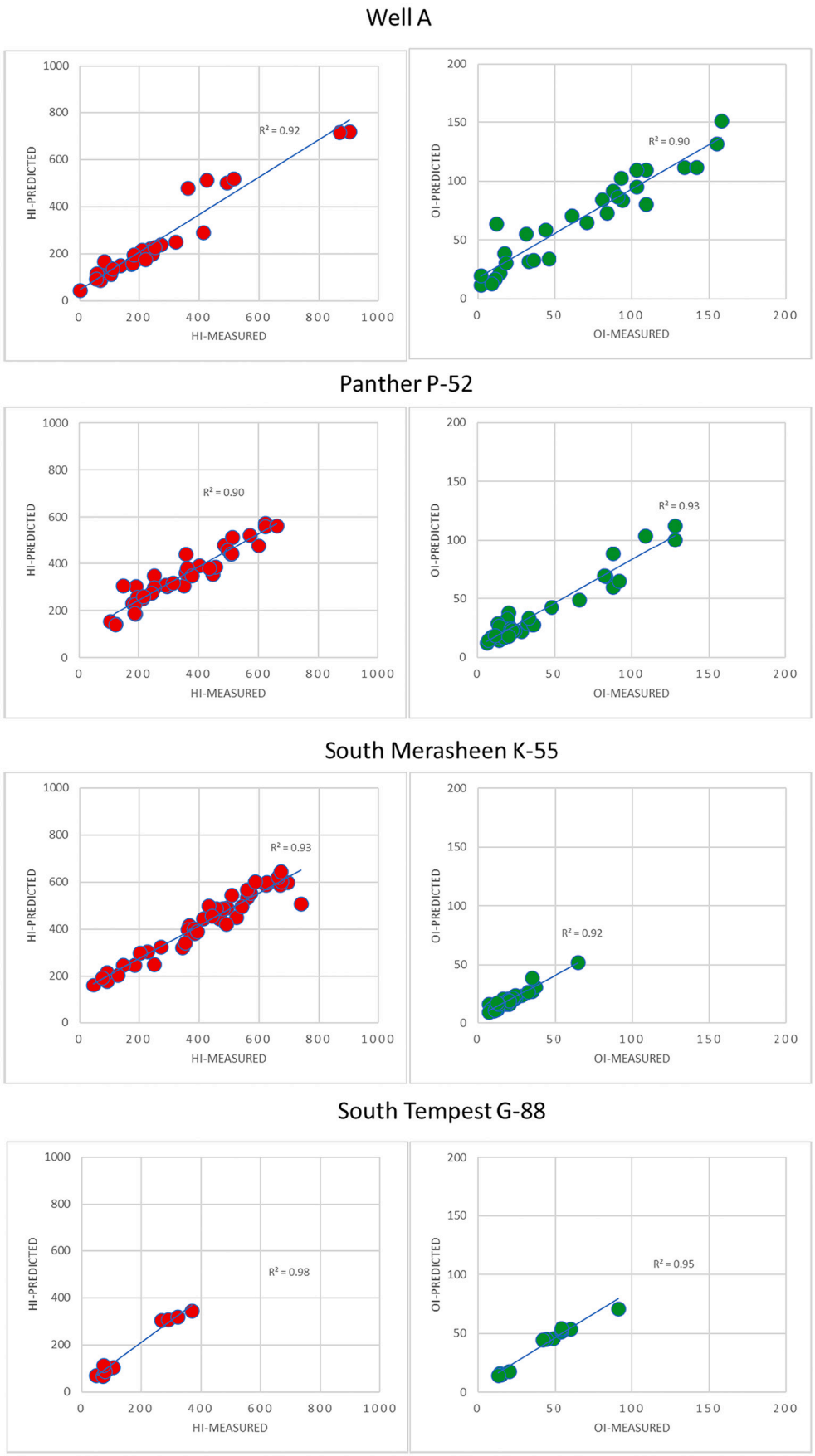


Fig. 9. Scatter plots of each well showing excellent correlation of predicted Random Forest data and programmed pyrolysis data. Note that some data are not normally distributed. Spearman's rank correlation was applied, and all data was found to have very high correlations.

Table 5
Error matrices for both HI and OI.

Well name	MAE		RMSE		R^2		Spearman rank correlation		# of samples
	HI	OI	HI	OI	HI	OI	HI	OI	
Well A	47.14	12.79	71.51	16.82	0.92	0.90	0.97	0.94	29
Panther P-52	47.78	8.65	60.41	11.51	0.90	0.93	0.89	0.83	37
S.Merasheen K-55	52.48	2.82	69.42	3.85	0.93	0.92	0.96	0.93	41
S.Tempest G-88	17.30	3.58	21.43	6.26	0.98	0.95	0.87	0.96	12

4.2.3. Transition ($n = 31$)

This sample grouping represents a transition from excellent hydrocarbon potential to poor (Fig. 7). Pyrolysis data for show HI ranging 401 to 599 (mg HC/g TOC), OI ranging 7.0 to 24.0 (mg CO_2/g TOC), and TOC ranging 1.38 to 8.62 (wt%). These samples exhibit laminations with abundant silt-sized siliciclastic gains with scattered OM in abundant calcareous and argillaceous matrix. Rare fossil fragments are also present. Point count data ($n = 5$) shows liptinite to still be the main OM present (68.0 to 97.0%). However, reworked vitrinite + inertinite (3.0 to 27.0%), and bituminite (0.0 to 17.0%) are showing more abundance than group above (Fig. 5c and d).

4.2.4. Poor hydrocarbon potential ($n = 75$)

This sample grouping shows HI ranging 0 to 393 (mg HC/g TOC), OI ranging 11.0 to 158.0 (mg CO_2/g TOC), and TOC ranging 0.28 to 5.08 (wt%; Fig. 7). These samples exhibit massive to laminated calcareous to silty/sandy mudrocks to siltstones to OM-poor lime mudrocks with very low hydrocarbon potential. A wide range of preserved OM are present in these samples. Point count data ($n = 13$) shows scattered liptinite (24.0 to 89.0%). An abundance of reworked vitrinite + inertinite (11.0 to 66.0%), and bituminite (0.0 to 14.0%; Fig. 5e and f).

4.2.5. Thermal maturity

The sample groupings above, being based solely on TOC, HI, and OI, ignores the thermal maturity of the OM. Measured VRo% on vitrinite macerals ranges 0.25 to 0.87%. Converting Tmax values to %VRo equivalent (Eq. 6; Jarvie, 2012) ranges 0.40 to 0.81% which is in good relation to the measured data (Table. 4). Caution should be used with this equation (Eq. 6) to convert Tmax to %VRo as there is no universal correlation for Tmax and %VRo and the equation likely differs from mudrock unit to mudrock unit globally (Yang and Horsfield, 2020).

$$\%VRo\ equivalent = (0.0180 \times \mathrm{Tmax}) - 7.16 \quad (6)$$

These data indicate immature to oil window thermal maturity. However, both of these thermal maturity methods ignore the contribution of algal OM as these maceral types cannot be measured using these techniques (Thompson-Rizer and Woods, 1987). Gordon et al., 2021 showed by integrating Fluorescence Red/Green Quotient (R/G Q) spectral data collected on a subsample set used in this study, measured only on preserved fluorescing algae, the %VRo equivalent ranges 0.48 to 0.61% indication that the thermal maturity has not yet reached the primary oil generation window. The variation of these measured values is due to the mixing and dilution of the higher reflecting and reworked vitrinite macerals, sourced from different depositional environments caused by variation in sea level (Gordon et al., 2021). Therefore, the R/G Q %VRo equivalent is likely a more accurate estimation of the thermal maturity in these samples.

4.3. RF model performance and prediction

Prior to the RF prediction model, the relationships between HI, OI, and the well log parameters were investigated using a series of scatter plots (Fig. 8). The results show no correlation between the HI, OI and the well log parameters. HI and OI were the only predicted attributes used in the RF model.

The RF modelling is predicting HI and OI with significant correlation coefficients that range from 0.90 to 0.98 and 0.90 to 0.95 R^2 respectively (Fig. 9). The MAE for HI and OI values range from 17.30 to 52.48 and 2.82 to 12.79, respectively. The RMSE for HI and OI range from 21.43 to 71.51 and 3.85 to 16.82, respectively. These values are relatively small given the overall large variation in HI and OI data. The Spearman's rank correlation for HI and OI range from 0.87 to 0.97 and 0.90 to 0.96, respectively (Table 5). The South Tempest G-88 has the lowest MAE and RMSE as the HI and OI variation is relatively small with HI ranging from 48 to 369 mg-HC/g TOC and OI ranging from 13 to 91 mg-CO_2/g TOC. Well A has the highest variation with HI ranging from 0 to 907 mg-HC/g TOC and OI ranging from 2 to 158 mg-CO_2/g TOC. This indicates the error values is likely only significant as the HI and/or OI values approach the applied cut-offs mentioned above for each of the three simplified groupings. Therefore, one must use caution as HI and OI values approach preconceived cut-off values. It should be noted that data overfitting can be an issue when using complex machine learning algorithms. However, since each tree is trained on a unique subsample of the data using bootstrapping techniques, RF models are robust and resistant to overfitting (Grimm et al., 2008; Handhal et al., 2020). A visual comparison of the final output of predicted HI and OI compared to the measured results can be seen in Fig. 10 for each well. A pseudo van Krevelen comparison of the measured vs. the predicted HI and OI for each well can be seen in Fig. 11. The HI prediction curves can predict the organic richness of each well where there were no samples taken. Similarly, the predicted OI graphs are emphasizing the depth where it is organic lean. These predictions further suggest organic richness is directly related to water depth and changes in depositional environment that affect the factors controlling the accumulation and preservation of organic matter.

5. Conclusions

This study tested the validity of using a Random Forest (RF) machine learning algorithm to predict hydrogen index (HI) and oxygen index (OI) using wireline logs. The RF model was trained using programmed pyrolysis data (Rock-Eval 6 method) and detailed organic petrology. The predictions of HI and OI from the RF model was evaluated using four statistical error matrices. The results showed the RF prediction performed very well with acceptable mean absolute error (MAE), root mean square error (RMSE), high correlation of determination (R^2) and high Spearman's rank correlations (R_2). This study showed that by using an integrated approach and using machine learning algorithms the prediction of important geochemical parameters from wireline logs can be satisfactorily achieved. Moreover, the results showed that samples with excellent hydrocarbon potential (high HI with low OI) are largely controlled by depositional environment. High hydrocarbon potential samples show an abundance of organic-rich lamalginites and filamentous alginite related to deeper offshore marine depositional environments. Low hydrocarbon potential samples (low HI and high OI) are related to the dilution by clastic terrigenous organic matter input due to proximity to deltaic sediment source. This is evident by the observed abundance of reworked vitrinite, and inertinite macerals scattered organic matter in the silty argillaceous matrix.

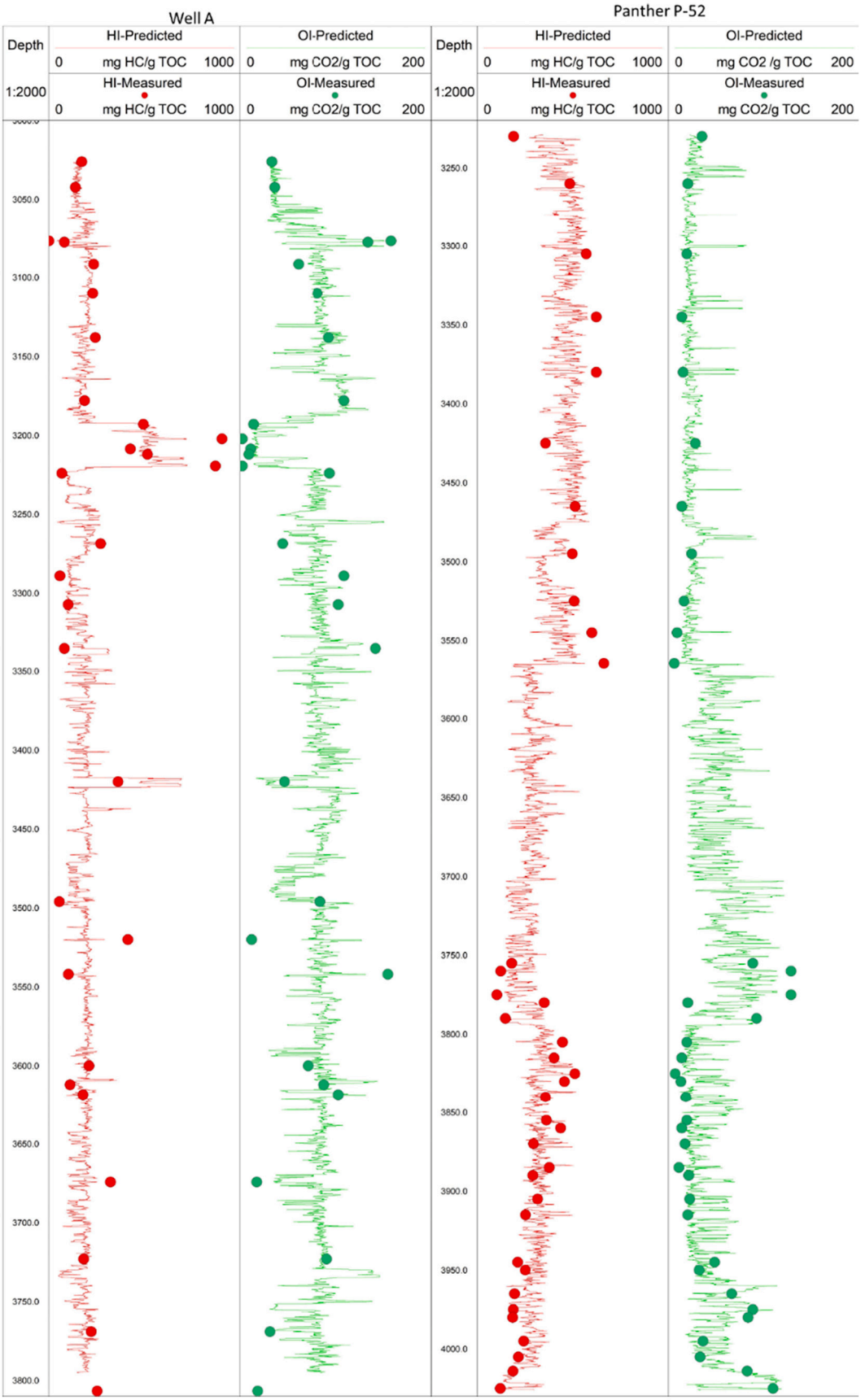


Fig. 10. Final Random Forest output of HI and OI prediction for Well A, Panther P-52, South Merasheen K-55 and South Tempest G-88 wells.

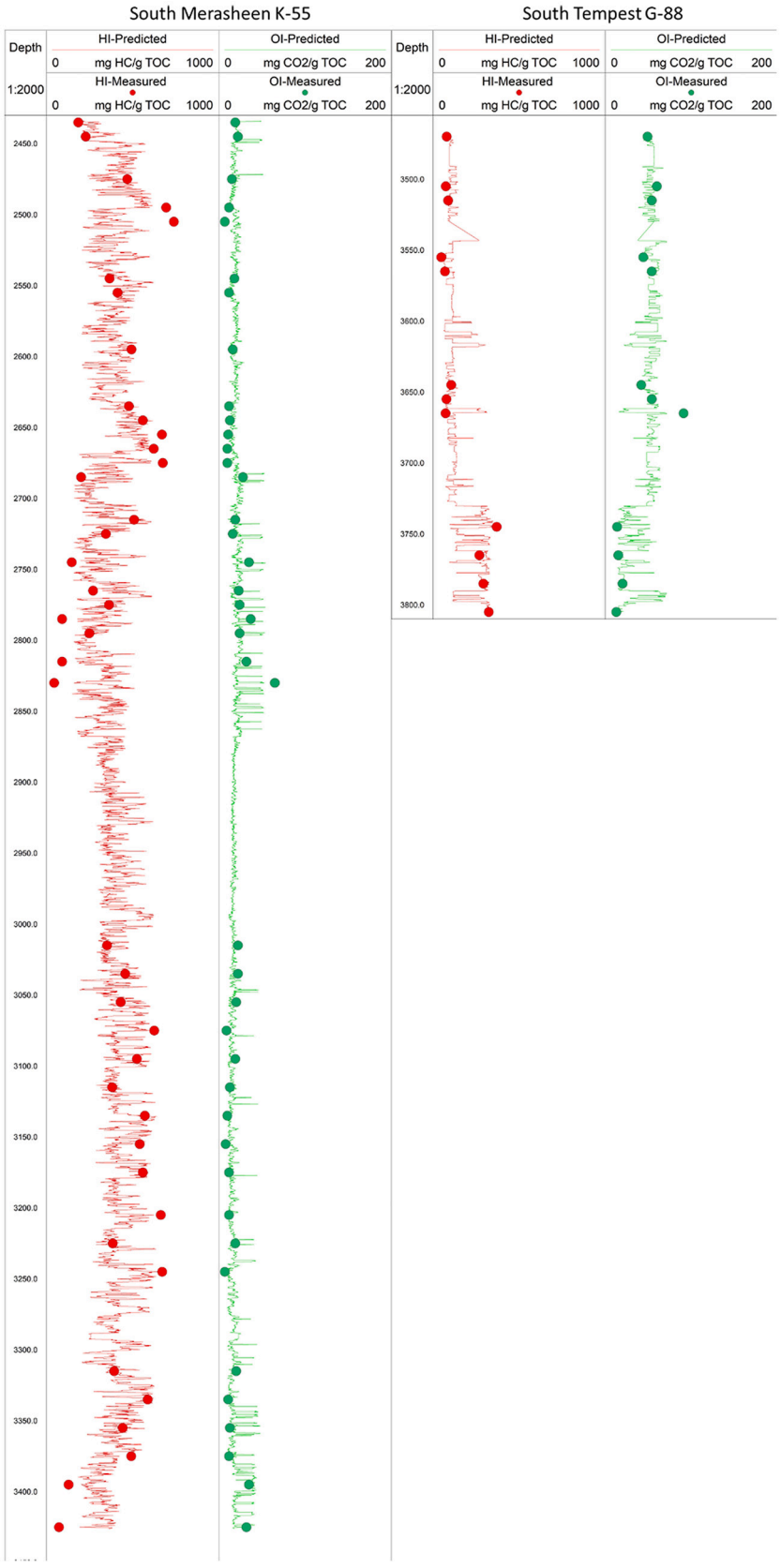


Fig. 10. (*continued*).

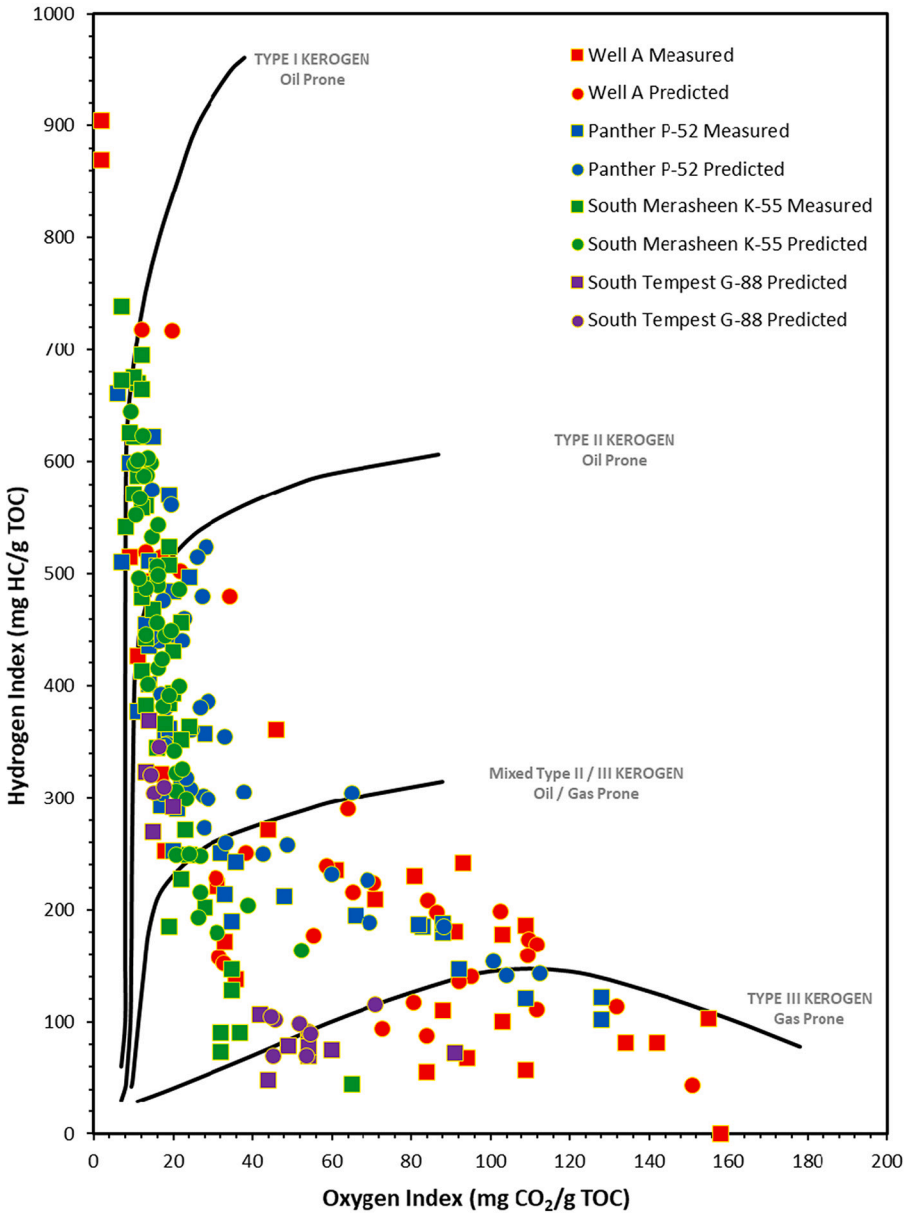


Fig. 11. Pseudo van Krevelen plot showing good correlation between measured HI and OI vs. Predicted HI vs. OI.

Contribution of authors

1) John B. Gordon (corresponding author)

- The conception and design of the study
- Analysis and interpretation of data (organic petrology, geochemistry)
- Drafting the article or revising it critically for important intellectual content

2) Hamed Sanei

- Interpretation of data (organic petrology, geochemistry)
- Revising the manuscript critically for important intellectual content

3) Per K. Pedersen

- Interpretation of data (depositional setting)
- Revising the manuscript critically for important intellectual content

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are extremely grateful to Dr. Omid H. Ardakani from the Geological Survey of Canada for the use of their lab facilities as well as Mun Tatt Yap and Jeff Unich from Cenovus Energy Inc. The manuscript improved greatly with suggestions supplied from two anonymous reviewers.

References

BeicipFranlab, 2015. Offshore Newfoundland & Labrador resource assessment Flemish Pass Area NL15_01EN. In: An integrated project for: Nalcor Energy – Oil and Gas Inc. Department of Natural Resources, Government of Newfoundland and Labrador.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogram. Remote Sens. 114, 24–31.
Bolandi, V., Kadkhodaie, A., Farzi, R., 2017. Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran. J. Petrol. Sci. Eng. 151, 224–234.
Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
Carvajal-Ortiz, H., Gentzis, T., 2015. Critical considerations when assessing hydrocarbon plays using Rock-Eval pyrolysis and organic petrology data: data quality revisited. Int. J. Coal Geol. 125 (part A), 113–122.
Cotterill, J.L., 1987. Well History Report Petro-Canada Lancaster G-70 (Petro-Canada Internal Report. Unpublished).
Creaney, S., Allison, B.H., 1987. An organic geochemical model of oil generation in the Avalon/Flemish Pass sub-basins, east coast Canada. Bull. Can. Petrol. Geol. 35 (1), 12–23.
Creaney, S., Passey, Q.R., 1993. Recurring patterns of total organic carbon and source rock quality within a sequence stratigraphic framework. AAPG Bull. 77 (3), 386–401.
Dembicki, H., 2009. Three common source rock evaluation errors made by geologists during prospect or play appraisals. AAPG Bull. 93, 341–356.
DeSilva, N.R., 1999. Sedimentary basins and petroleum systems offshore Newfoundland and Labrador. In: Fleet, A.J., Boldy, S.A.R. (Eds.), Petroleum Geology of Northwest Europe: Proceedings of the Fifth Conference. The Geological Society, London, pp. 501–515.
Enachescu, M.E., 2005. Offshore Newfoundland and Labrador: an emerging energy powerhouse. In: Offshore Technology Conference Paper #17570, Houston, Tx (8 pp.).
Enachescu, M., 2012. Petroleum Exploration Opportunities in the Flemish Pass Basin. Call for Bids NL12-02, Parcel 1. http://www.nr.gov.nl.ca/nr/invest/call_bids_petro_exploration_enachescu%20.pdf.
Enachescu, M.E., Hogg, J.R., Fowler, M., Brown, D.E., Atkinson, I., 2010. Late Jurassic Source Rock Super-highway on Conjugate Margins of the North and Central Atlantic (Offshore East Coast Canada, Ireland, Portugal, Spain and Morocco). CM 2010-Abstracts, 2.
Espitalie, J., Madec, M., Tissot, B., Mennig, J.J., Leplat, P., 1977. Source rock characterization method for petroleum exploration. In: Offshore Technology Conference Estimation of Total Organic Carbon From Well Logs and Seismic Sections Via Neural Network and Ant Colony Optimization Approach: A Case Study From the Mansuri Oil Field, SW Iran, 7. Geopersia, pp. 255–266, 2017.
Fowler, M.G., McAlpine, K.D., 1995. The Egret Member, a Prolific Kimmeridgian source rock from Offshore Eastern Canada. In: Katz, B.J. (Ed.), Petroleum Source Rocks. Casebooks in Earth Sciences. Springer, Berlin, Heidelber.
Fowler, M.G., Snowdon, L.R., Stewart, K.R., McAlpine, K.D., 1990. Rock-Eval/TOC data from nine wells located offshore. In: Newfoundland Geol. Surv. Can. Open File Rep. 2271, p. 72.
Fowler, M.G., Snowdon, L.R., Stewart, K.R., McAlpine, K.D., 1991. Rock-Eval/TOC data from five wells located within Jeanne d'Arc Basin, offshore. In: Newfoundland GeoL Surv. Can.Open File Rep. 2392, p. 41.
Fowler, M.G., Obermajer, M., Achal, S., Milovic, M., 2007. Results of Geochemical Analyses of an Oil Sample From Mizzen L-11 Well, Flemish Pass, Offshore Eastern Canada. Open-File Report. Geological Survey of Canada.
Gordon, J.B., Sanei, H., Ardakani, O.H., Pedersen, P.K., 2021. Effect of sediment source on source rock hydrocarbon potential; an example from the Kimmeridgian and Tithonian-aged source rocks of the central ridge, off-shore Newfoundland, Canada. Mar. Petrol. Geol. 127.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis. Geoderma 146 (1–2), 102–113.
Hackley, P., Zhang, T., Jubb, A., Valentine, B., Dulong, F., Hatcherian, J., 2020. Organic petrography of Leonardian (Wolfcamp A) mudrocks and carbonates, Midland Basin, Texas: the fate of oil-prone sedimentary organic matter in the oil window. Mar. Petrol. Geol. 112.
Handhal, A.M., Al-Abadi, A.M., Chafeet, H.E., Ismail, M.J., 2020. Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. Mar. Petrol. Geol. 116.
Huang, Z., 1994. Predicted and measured petrophysical and geochemical characteristics of the Egret Member oil source rock, Jeanne d'Arc Basin, Offshore Eastern Canada. Mar. Pet. Geol. 11, 294–306.
ICCP, 1998. The new vitrinite classification (ICCP System 1994). Fuel 77, 349–358.
ICCP, 2001. The new inertinite classification (ICCP System 1994). Fuel 80, 459–471.
Jarvie, D.M., 2012. Shale resource systems for oil and gas: Part 1 – shale oil resource systems. In: Breyer, J. (Ed.), Shale Reservoirs – Giant Resources for the 21st Century: AAPG Memoir, vol. v. 97, pp. 1–19.
Keykhay-Hosseinpoor, M., Kohsary, A.H., Hossein-Morshedy, A., Porwal, A., 2020. A machine learning-based approach to exploration targeting of porphyry Cu-Au deposits in the Dehsalm district, eastern Iran. Ore Geol. Rev. 116.
Lafargue, E., Marquis, F., Pillot, D., 1998. Rock-Eval 6 applications in hydrocarbon exploration, production and soil contamination studies. Revue De L'institut Francais Du Pétrole 53 (4), 421–437.
Omura, A., Hoyanagi, K., 2004. Relationships between composition of organic matter, depositional environments, and sea-level changes in backarc basins, Central Japan. J. Sediment. Res. 74, 620–630.
Passey, Q.R., Creaney, S., Kulla, J.B., Moretti, F.J., Stroud, J.D., 1990. A practical model for organic richness from porosity and resistivity logs. AAPG Bull. 74 (12), 1777–1794.
Passey, Q.R., Bohacs, K.M., Esch, W.L., Klimentidis, R., Sinha, S., 2012. My source rock is now my reservoir - geologic and petrophysical characterization of shale-gas reservoirs. Search Discov., 80231
Peters, K., Xia, X., Pomerantz, A., Mullins, O., 2015. Geochemistry applied to evaluation of unconventional resources. In: Unconventional Oil and Gas Resources Handbook: Evaluation and Development. Elsevier, pp. 71–126.
Pickel, W., Kus, J., Flores, D., Kalaitzidis, S., Christanis, K., Cardott, B.J., Misz-Kennan, M., Rodrigues, S., Hentschel, A., Hamor-Vido, M., Crosdale, P., Wagner, N., 2017. Classification of liptinite – ICCP system 1994. Int. J. Coal Geol. 169.
Probst, P., Wright, M., Boulesteix, A.l., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. 9 (3).
Raine, R., 2006. Petrographical Analyses and Depositional Environmental Interpretation of Laminated Mudrocks From Well L-76Z, Bay du Nord, Offshore Newfoundland (Internal report prepared for Statoil Canada Limited.).
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71.
Sanei, H., 2020. Genesis of solid bitumen. Sci. Rep. 10, 15595.
Sun, T., Chen, F., Zhong, L., Liu, W., Wang, Y., 2019. GIS-based mineral prospectivity mapping using machine learning methods: a case study from Tongling ore district, eastern China. Ore Geol. Rev. 109.
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inform. Comput. Sci. 43, 1947–1958.
Swift, J.H., Williams, J.A., 1980. Petroleum source rocks, grand banks area. In: Miall, A. D. (Ed.), Facts and Principles of World Petroleum Occurrence: Canadian Society of Petroleum Geologists Memoir, 6, pp. 567–587.
Thompson-Rizer, C.L., Woods, R.A., 1987. Microspectrofluorescence measurements of coals and petroleum source rocks. Int. J. Coal Geol. 7, 85–104.
Tissot, B.P., Durand, B., Espitalié, J., Combaz, A., 1974. Influence of nature and diagenesis of organic matter in formation of petroleum. AAPG Bull. 58, 499–506.
Wang, H., Wu, W., Chen, T., Dong, W., Wang, G., 2019a. An improved neural network for TOC, S1 and S2 estimation based on conventional well logs. J. Petrol. Sci. Eng. 176.
Wang, Z., Dong, Y., Zuo, R., 2019b. Mapping geochemical anomalies related to Fe–polymetallic mineralization using the maximum margin metric learning method. Ore Geol. Rev. 107, 258–265.
Yang, S., Horsfield, B., 2020. Critical review of the uncertainty of tmax in revealing the thermal maturity of organic matter in sedimentary rocks. Int. J. Coal Geol. 225, 103500.