Research article

PG
Petroleum Geoscience

https://doi.org/10.1144/petgeo2022-071 | Vol. 29 | 2023 | petgeo2022-071

Predicting oil field performance using machine learning programming: a comparative case study from the UK continental shelf

Ukari Osah*[*] and John Howell

Department of Geology, University of Aberdeen, Aberdeen AB24 3UE, UK
UO, 0000-0003-3265-7763; JH, 0000-0003-3818-8048
* Correspondence: ukari.osah@abdn.ac.uk

Abstract: Predicting the performance of a subsurface oil field is a large, multivariant problem. Production is controlled and influenced by a wide array of geological and engineering parameters which overlap and interact in ways that are difficult to unravel in a manner that can be predictive. Supervised machine learning is a statistical approach which uses empirical learnings from a training dataset to create models and make predictions about future outcomes. The goal of this study is to test a number of supervised machine learning methods on a dataset of oil fields from the United Kingdom continental shelf (UKCS), in order to assess whether, (a) it is possible to predict future oil field performance and (b), which methods are the most effective. The study is based on a dataset of 60 fields with 5 controlling parameters, (gross depositional environment, average permeability, net-to-gross, gas–oil ratio and total number of wells) and 2 outcome parameters (recovery factor and maximum field rate) for each. The choice of controlling parameters was based on a PCA of a larger dataset from a wider project database. Five different machine learning algorithms were tested. These include linear regression, robust linear regression, linear kernel support vector regression, cubic kernel support vector regression and boosted trees regression. Overall, 83% of the data was used as a training dataset while 17% was used to test the predictability of the algorithms. Results were compared using R-Squared, Mean Square Error, Root Mean Square Error and Mean Absolute Error. Graphs of predicted responses v. true (actual) responses are also shown to give a visual illustration of model performance. Results of this analysis show that certain methods perform better than others, depending on the outcome variable in question (recovery factor or maximum field rate). The best method for both outcome variables was the support vector regression, where, depending on the kernel function applied, a reliable level of predictability with low error rates were achieved. This demonstrates a strong potential for statistics-based prediction models of reservoir performance.

Thematic collection: This article is part of the Digitally enabled geoscience workflows: unlocking the power of our data collection available at https://www.lyellcollection.org/topic/collections/digitally-enabled-geoscience-workflows

Received 13 September 2022; **revised** 27 December 2022; **accepted** 3 January 2023

The efficiency of the hydrocarbon extraction process is largely dependent on a host of interconnected factors, both intrinsic and imposed. The goal of this study is to investigate the ability of machine learning algorithms to produce a predictive tool which can be applied to other datasets.

This paper forms part of larger project with a database that comprises 424 fields on the UKCS. A subset of that database was analysed using methods of feature selection including principal component analysis (PCA) and best subset regression to determine which variables were critical to predicting reservoir performance. In this paper, those variables have been used to condition a number of machine learning algorithms to determine which are the most effective at predicting future field performance.

Variables that control reservoir performance have been subdivided into geological, PVT (fluids and reservoir conditions) and engineering. A number of metrics that record reservoir performance were identified, and for the purpose of this study two of these response variables were selected (recovery factor and maximum field production rate). The original project database included information from 424 fields subsampled into smaller subsets for PCA and regression analysis. A further subsampling has been undertaken here for this analysis in a subset for Machine Learning. This subset of the database includes information about 5 control and 2 outcome variables from 60 fields. These fields and variables are an outcome of both the PCA and a best subsets regression testing. The data were z-score standardized and used to test five different machine learning methods.

The study area (shown in Fig. 1) was chosen for its wealth of exploration and production data accumulated over fifty plus years as a hydrocarbon producing region. Production data were obtained from the UK Oil and Gas Authority, (www.ogauthority.co.uk) and geological parameters were compiled from a variety of published sources. A comprehensive list of references and data sources and more detailed discussions on study area, data distribution and petroleum system geology are provided in a separate publication discussing the database building process and the spatio-temporal distribution of that data.

Study area

Sixty oil fields were selected for this study from the wider database of 424 oil, gas and condensate fields. These fields were randomly chosen based on fluid phase (oil) and filtered for completeness of data and to remove outliers and is representative of the region's oil fields spanning 4 separate basins. See Table 1 for a list of fields and Figure 1 for a map showing the location and spatial distribution of the fields used in this machine learning exercise.

The fields used in this study are from north of the Mid-North Sea High. Half are located in the Northern North Sea basin, a quarter in the Central North Sea basin and another quarter in the Moray Firth basin. 26 of these fields are strictly shallow marine, 25 are deep marine, 2 are continental and the remaining 7 contain a mix of gross depositional environments (including Chanter, Claymore,

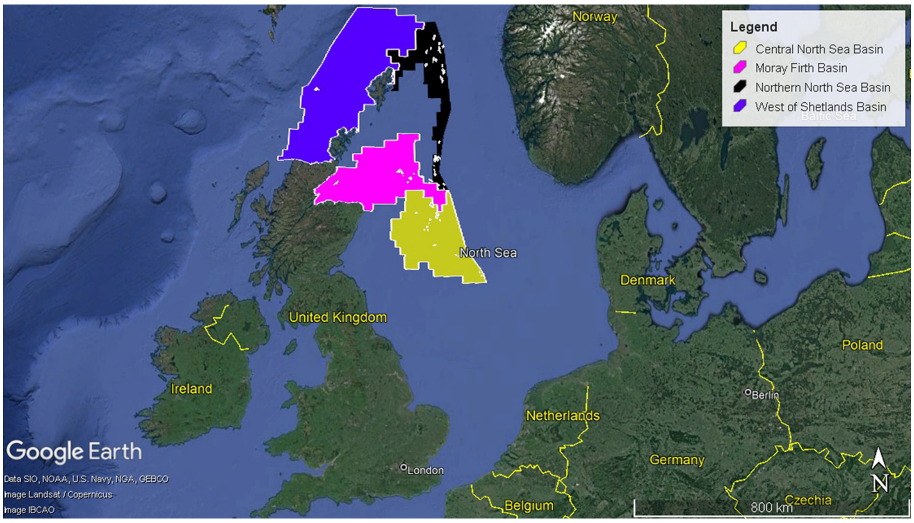

Fig. 1. Map of UKCS showing oil fields used in machine learning with sedimentary basins highlighted. (Created with Google Earth Pro).

Crawford, Dunbar, Fulmar, Highlander and Maureen). In instances with multiple depositional environments the primary reservoir accounting for over 70%–80% of in place and produced volumes was used for the GDE classification. About 35 of these have further sedimentological data that was not used directly in this machine learning study (e.g. diagenetic impact, stratigraphic heterogeneity, etc.) which were recorded as having low to moderate intensity. Trapping mechanisms were mostly structural at reservoir depths between 1335 and 3980 m. The hydrocarbons were light crudes (mean of 38° API) in reservoirs with good porosities averaging >20%. Reservoirs are predominantly Jurassic in age with a few Triassic, Paleocene and other age.

This supervised learning experiment utilizes 5 predictor variables including gross depositional environment (GDE), average permeability, net-to-gross (NTG), gas–oil ratio (GOR) and total number of wells. These parameters were chosen from a wider selection of 27 predictor variables based on PCA which were then put through best-of-subset testing to assess minimum number of variables needed for prediction and suitable permutations (combinations) to achieve desired results. All variables applied here ranked among those that were found to control 83% of the correlation in the predictor variables of the PCA.

A summary of the feature selection process includes the following steps

- Classification of input data into three groups (categorical, ordinal and numerical variables); where categorical refers to descriptive or qualitative data points such as gross depositional environment; ordinal refers to numerical data with no order or magnitude such as structural complexity; and numerical refers to data that are number and connote a change in intensity with ascension such as permeability.
- Division of database into subsets of differing sample sizes (38 v. 136 oil fields) with overlapping variables to determine the impact of sample size on results. Results were seen to be consistent across sample sizes.
- Preparation of data for statistical analyses including label encoding of non-numerical data and standardization of numerical data.
- Principal component analysis (PCA) for feature selection to determine how variables interact with each other. PCA works by projecting numbered data into lower dimensions called principal components for the purpose of finding the most succinct/effective expression of all input parameters using a limited number of principal components (Lever *et al.* 2017). The main output of the PCA is expressed as a table of eigenvalues from which to determine the suitable number of principal components based on a cut-off at either the first principal component to exceed the 80% cumulative proportion threshold (Jolliffe's rule) or principal components with eigenvalues greater than 1 (Kaiser criterion) (Jolliffe 2002). Selected variables for each principal component are determined based on the magnitude of eigenvectors.

The Eigenanalysis matrix (Table 2) and the table of results for selected principal components (Table 3) are shown below. This gives a sense of the wider selection of variables and the narrowing-down enabled by the PCA exercise.

It is not the purpose of this selection of five variables to be made a definitive recommendation on what parameters should be used to predict recovery factor or any other oil production performance measures. This is simply one permutation out of many, based on a

Table 1. *List of fields used in machine learning exercise*

Field list – machine learning					
ALWYN NORTH	CLAYMORE	EIDER	GANNET G	MAGNUS	ROB ROY
ANDREW	CLYDE	EMERALD	GLAMIS	MAUREEN	SCAPA
ARBROATH	CRAWFORD	FIFE	HAMISH	MILLER	SCOTT
ARKWRIGHT	CURLEW	FLORA	HEATHER	MOIRA	STIRLING
BALMORAL	CYRUS	FOINAVEN	HIGHLANDER	MONTROSE	STRATHSPEY
BEATRICE	DEVERON	FULMAR	HUTTON	MURCHISON	TERN
BIRCH	DON NORTHEAST	GANNET C	IVANHOE	NELSON	THELMA
BRAE CENTRAL	DON SOUTHWEST	GANNET D	KINGFISHER	NORTHWEST HUTTON	THISTLE
BRAE SOUTH	DUNBAR	GANNET E	KITTIWAKE	OSPREY	TIFFANY
CHANTER	DUNLIN	GANNET F	MACCULLOCH	PETRONELLA	TONI

limited dataset with a decision informed by the work carried out as part of the larger project.

A brief discussion of the impact of these predictor variables on reservoir performance is as follows:

- Gross Depositional Environment (GDE): the GDE of reservoir sediments imparts well understood characteristics that play a major role in controlling the level of productivity of fluids from within. The depositional environment controls the architecture and geometry of both reservoir bodies and baffles. It also affects the textural properties and the mineralogy of the reservoirs (Lorenz *et al.* 1989; Ingles and Anadon 1991; Reinson 1991; Hartmann and Beaumont 1999; Zhang *et al.* 2008; Armitage *et al.* 2010; Lai *et al.* 2015; Wang *et al.* 2018; Ärlebrand *et al.* 2021), pore-water chemistry (Shaw *et al.* 1990; Hartmann and Beaumont 1999; Toevs *et al.* 2008) and reservoir geometry/structural control (Reinson 1991; Mode *et al.* 2017; Levell 2021). This variable is categorical with three classes of Continental, Shallow Marine and Deep Marine. This variable was also selected because it ranked highly in the PCA, being one of the variables that accounted for 47% of the correlation in the predictor data.
- Net-to-Gross (NTG): refers to the proportion of the gross reservoir volume that can hold and deliver hydrocarbons. As a general rule, low net-to-gross reservoirs are associated with poor recovery factors (e.g. Richards and Bowman 1998) but this is not an explicit relationship as it does not capture geometry or architecture of reservoirs or baffles. Net-to-gross is a ratio between 0–1. Within the current dataset net-to-gross ranged from 0.35 to 1 with an average value of 0.73. This is one of the variables that accounted for 83% of correlation in the predictor data.
- Average Permeability: permeability is the capacity of the reservoir to transmit its fluid contents through the pore network and internal fractures and fissures. This is a key component of reservoir quality assessment metrics and has great effect on the performance of the reservoir (Gunter *et al.* 1997). Discounting any flaws in production strategy, permeability and its partner index (porosity) provide a fair assessment of potential production efficiency. This metric was recorded in milliDarcy (mD) and average permeabilities were between <1 to 2000 mD, with a mean value over 500mD. This parameter ranked highly in the PCA being one of the parameters that accounted for 55% of the correlation in the data.
- Gas–Oil Ratio (GOR): this refers to the amount of gas in solution relative to a unit volume of oil at reservoir conditions. GOR as a predictive element has also played a role in previous studies of reservoir performance prediction including material balance equations. Ahmed and McKinney (2005) and Ahmed and Meehan (2012) dissect the intricacies of the topic in greater detail including equations for determining GOR as well as the relevance of

Table 2. *Eigenanalysis matrix showing eigenvalue and eigenvector distributions for the larger dataset*

	Eigenanalysis of the correlation matrix								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalue	**6.366**	**3.559**	**2.687**	**2.331**	**1.793**	**1.633**	**1.479**	**1.331**	**1.084**
Proportion	**0.236**	**0.132**	**0.100**	**0.086**	**0.066**	**0.060**	**0.055**	**0.049**	**0.040**
Cumulative	**0.236**	**0.368**	**0.467**	**0.553**	**0.620**	**0.680**	**0.735**	**0.784**	**0.825**
Eigenvectors									
Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
Gross. Dep. Environment	**0.212**	**0.077**	**−0.358**	**0.101**	**0.224**	**0.096**	**−0.006**	**0.077**	**−0.061**
Reservoir Depth (m)	**0.234**	**0.229**	**−0.221**	**−0.130**	**0.195**	**−0.234**	**−0.013**	**0.094**	**0.249**
Avg. Porosity (%)	**−0.155**	**0.282**	**−0.125**	**0.344**	**−0.163**	**0.198**	**0.065**	**−0.086**	**−0.088**
Avg. Permeability (mD)	**−0.091**	**0.126**	**−0.219**	**0.365**	**−0.295**	**0.279**	**−0.053**	**−0.069**	**0.134**
Pressure (bar)	**0.251**	**0.267**	**−0.131**	**−0.144**	**−0.130**	**−0.003**	**−0.068**	**−0.182**	**0.261**
Temperature (°C)	**0.140**	**0.411**	**−0.103**	**−0.123**	**0.122**	**−0.178**	**0.115**	**0.027**	**0.086**
Net:Gross	**−0.242**	**−0.165**	**−0.100**	**0.100**	**−0.181**	**−0.066**	**−0.219**	**0.051**	**0.443**
Fault Compartments	**0.244**	**−0.098**	**−0.137**	**−0.173**	**−0.130**	**0.028**	**−0.105**	**−0.393**	**−0.135**
Structural Complexity	**0.264**	**−0.163**	**−0.134**	**−0.197**	**−0.093**	**0.178**	**−0.114**	**−0.327**	**−0.064**
No. of Fault populations	**0.230**	**−0.073**	**−0.259**	**−0.111**	**0.084**	**0.164**	**−0.107**	**−0.059**	**0.128**
API (°)	**−0.086**	**−0.228**	**0.050**	**0.105**	**0.265**	**−0.389**	**0.210**	**−0.130**	**0.324**
Field Area (km^2)	**0.211**	**0.051**	**0.338**	**0.275**	**0.175**	**−0.033**	**−0.180**	**0.036**	**−0.050**
Bulk Rock Volume (10^6 m^3)	**0.295**	**−0.010**	**0.242**	**0.067**	**0.035**	**−0.045**	**−0.267**	**−0.063**	**0.081**
Well Density (wells/km^3)	**−0.113**	**−0.186**	**−0.324**	**0.106**	**0.084**	**0.118**	**0.056**	**0.160**	**0.085**
No. of Production wells	**0.291**	**−0.139**	**−0.023**	**0.270**	**−0.152**	**−0.078**	**0.221**	**0.051**	**−0.077**
No. of Injection wells	**0.277**	**−0.147**	**0.040**	**0.290**	**−0.039**	**−0.020**	**0.239**	**0.036**	**−0.040**
Average GOR (m^3 m$^{−3}$)	**−0.123**	**−0.118**	**0.113**	**0.082**	**−0.073**	**−0.146**	**0.015**	**−0.657**	**0.008**
Max. Thickness (m)	**0.037**	**0.026**	**0.173**	**−0.178**	**−0.505**	**−0.121**	**0.177**	**0.101**	**0.418**
Water Saturation (%)	**−0.045**	**0.370**	**0.214**	**−0.009**	**−0.080**	**−0.041**	**0.294**	**−0.198**	**−0.025**
Production Strategy	**0.136**	**−0.034**	**0.108**	**0.008**	**0.123**	**0.421**	**0.102**	**0.038**	**0.445**
Trap Type	**0.007**	**−0.064**	**0.187**	**−0.161**	**0.050**	**0.500**	**0.351**	**−0.058**	**0.081**
Diagenetic impact	**−0.020**	**−0.058**	**0.081**	**−0.009**	**0.425**	**0.114**	**0.326**	**−0.166**	**0.121**
Paleoclimate	**−0.044**	**0.349**	**−0.199**	**0.259**	**0.060**	**−0.116**	**0.146**	**−0.198**	**0.052**
Stratigraphic Heterogeneity	**0.209**	**0.089**	**0.109**	**−0.208**	**−0.252**	**−0.033**	**0.191**	**0.255**	**−0.180**
Prod. well Spacing (km^2/well)	**−0.062**	**0.308**	**0.307**	**0.048**	**0.142**	**0.229**	**−0.250**	**0.002**	**0.001**
Total No. Wells	**0.297**	**−0.147**	**−0.006**	**0.291**	**−0.126**	**−0.065**	**0.243**	**0.051**	**−0.062**
OIP (Mill. Sm^3)	**0.239**	**0.005**	**0.218**	**0.270**	**0.012**	**−0.004**	**−0.313**	**0.027**	**0.190**

Eigenvectors for selected variables per principal component are highlighted in green.

Table 3. *Table of results for selected variables from PCA for larger dataset*

Principal components	Key variables
PC 1–24%	**Total Number of Wells** **Bulk Rock Volume (10^6 m^3)**
+ PC 2–37%	**Temperature (°C)**
+ PC 3–47%	**Gross. Dep. Environment** **Field Area (km^2)**
+ PC 4–55%	**Avg. Permeability (mD)** **Avg. Porosity (%)**
+ PC 5–62%	**Max. Thickness (m)**
+ PC 6–68%	**Trap Type**
+ PC 7–74%	**Diagenetic impact**
+ PC 8–78%	**Average GOR (m^3 m^{-3})**
+ PC 9–83%	**Production Strategy** **Net:Gross**

Cumulative Percentages of variation accounted for by principal components is indicated along with key variables of each principal component**.**

GOR in predicting reservoir performance. Busahmin and Maini (2010) also discuss how GOR affects recovery factor and production rate in the context of heavy oil reservoirs, observing a decrease in oil recovery with increasing GOR. The unit of measurement applied here is standard cubic meters per standard cubic meter (m^3 m^{-3}). GOR values were between 15 to >500 m^3 m^{-3} with a mean of >100 m^3 m^{-3}. This parameter ranked highly in the PCA, being one of the variables that contribute to 78% of the correlation in the data.

- Total Number of Wells: here we account for all well bores (both producers and injectors) on the field. As a key element of the production process, well related parameters greatly affect overall field performance (Gurbanov *et al.* 2016). In relation to field size, this parameter factors in well spacing and well density while also being dependent on production strategy (primary, secondary or tertiary) and chosen drive mechanism. Total number of wells provides a singular compound measure of external forces of extraction (production wells) and input (injection wells) at play on the reservoir. Total Number of wells for our experimental dataset lies between 1 and 77 with an average of 18 wells. This parameter was chosen because it was one of the variables that accounts for the top 24% of correlation in the control variable data.

For response variables, the two parameters chosen (Recovery Factor and Maximum Field Production Rate) give a measure of reservoir performance. These were selected from among other relevant outcomes with completeness/abundance of data being a consideration. Two parameters were assessed to gauge consistency of results across models and to test if model results vary depending on response variable, given the same predictors.

- Recovery Factor (RF): this is the percentage of in-place volumes of hydrocarbon which is producible as per implicit technicalities (including whether primary or enhanced recovery techniques are applied) or recovered as at the end of field life. Values recorded for this project were either forecasted as indicated in existing literature or are coincidental with present realized recovery at cease-of-production. Recovery factors ranged from 6% to 77% with an average value of 40%.
- Maximum Field Production Rate (MFPR): this is recorded in thousands of barrels of oil per day (mbpd) and indicates the ceiling of the achievable hydrocarbon extraction rate, recorded over production time span, through flow testing or during production. These rates are capped by a variety of physical factors and field planning decisions. As fields in this experiment are offshore fields, maximum flow rates are generally on the higher side, as profitability in offshore fields require higher production rates (Dake 1994). Larue and Friedmann (2005) suggest that flow from a reservoir is mostly influenced by the reservoir architecture which is related to GDE. Values for this metric are between 2 mbpd and 300 mbpd with a mean value of 58 mbpd.

Table 4 provides a summary of mean and range of values for each predictor and outcome variable.

Previous work

Mustafiz and Islam (2008) suggested that there are three main types of analysis that can be used for reservoir performance prediction.

- The Analogical Approach: a comparative and inferential assessment of reservoir performance hinged on similarities in characteristics between mature and early–stage zones or pools. This approach can be strictly qualitative or employ quantitative measures in the form of empirical statistics to observe correlations and approximate production; for example, as discussed in Meehan (2011) where analogues of fractured reservoirs were compared for performance.
- The Experimental Approach: here PVT and other properties are measured in lab models and observed results are scaled to the level of the actual reservoirs (Manzir *et al.* 2015).
- The Mathematical Approach: these methods apply mathematical equations to predict performance. Ertekin *et al.* (2001) gives a comprehensive description of mathematical methods including material balance equations, decline curves, statistical and analytical methods. Okotie and Ikporo (2018) also discuss material balance for performance prediction.

Machine learning combines the analogical and the mathematical approach. Here statistical equations are produced using requisite amounts of data samples with established independent-dependent (control-response) multivariate pairings. Derived equations are then applied to control variables to predict responses. Ertekin and Sun (2019), Pandey *et al.* (2020) and Sircar *et al.* (2021) also give broad and up to date overviews on the concepts behind the application of

Table 4. *Table of variables and data ranges for the predictor and outcome variables used in machine learning*

	Variables	Min	Max	Mean
Predictor	*Permeability (mD)*	0.68	2000.00	545.55
	Net-to-Gross	0.35	1.00	0.73
	Gas–Oil Ratio (m^3 m^{-3})	14.60	511.30	137.78
	Total Number of Wells	1.00	77.00	18.03
Outcome	*Recovery Factor (%)*	6.00	77.00	39.60
	Maximum Field Production Rate (kbpd)	2.61	300.51	58.27

machine learning in forward and inverse reservoir performance and reservoir quality modelling, although not specifically focusing on any singular unique case studies.

A recent case study application of machine learning in hydrocarbon reservoir performance prediction includes Niu *et al.* (2021) where data from 172 gas wells from a single producing block and a selection of 19 engineering and geological variables were compiled. Following feature selection, 8 variables were chosen and used to create an ultimate recovery prediction model based on multiple regression.

Other examples of case study applications of various artificial intelligence and machine learning techniques (such as genetic algorithms, random forest, artificial neural networks and others) over the last decade include Al-Fattah and Startzman (2001); Mirzaei-Paiaman and Salavati (2012); Amirian *et al.* (2013); Chithra Chakra *et al.* (2013*a*); Chithra Chakra *et al.* (2013*b*); Li *et al.* (2013); Ahmadi *et al.* (2015); Choubineh *et al.* (2017); Ghahfarokhi *et al.* (2018); Bhattacharya *et al.* (2019); Ghorbani *et al.* (2019); Aliyuda *et al.* (2020); Liu *et al.* (2020); Al-Jifri *et al.* (2021); and Han and Kwon (2021); Bhattacharyya and Vyas (2022*a*); and Bhattacharyya and Vyas (2022*b*).

Approach

For this project, 5 different statistical models based on 3 modelling techniques were developed using machine learning software – MATLAB R2019b Update 5 (9.7.0. 1319299). This was done to assess consistency in results and check potential biases that may present from any one chosen methodology. These modelling techniques broadly include:

- Linear regression
- Support vector regression
- Boosted trees regression

An overview of the technicalities of each method is provided below.

Machine learning is the development and implementation of algorithms that improve automatically with experience. A widely used description is provided by Mitchell (1997), defining it as software being able to learn from experience (E) in application to a specific set of tasks (T) and indicators of performance (P) ‘if its performance at tasks in T, as measured by P, improves with experience E’.

Various texts list a plethora of approaches to machine learning for different purposes, including supervised, unsupervised, semi-supervised, reinforcement, self, feature, sparse dictionary, anomaly detection, robot learning, association rules, etc.

For this experiment, supervised machine learning is applied. This form of machine learning model is trained with both input and output data as opposed to unsupervised learning where the model is trained to identify clusters and classes with no outcome variable provided for training (Russell and Norvig 2010).

In supervised machine learning, models are trained with the training dataset, where input and output variable pairings are complete for model fitting. Prior to model fitting, a method of model validation is selected. This effectively amounts to a portion of the data used to assess the fit of the model. Options for validation typically include k-fold cross validation, hold-out validation or bootstrap (Kohavi 1995). For our purposes, holdout validation was chosen. In this method of validation, a larger percentage of the data is used to train the model, typically 66.6%; while the remaining 33.3% is used for validation. In this case the split was 83% train and 17% validate. This method was thought best because the number of observations available (60) and the number of independent variables (5) were enough to train a single iteration, in any given instance of the model, to acceptable standards of observations per variable, applying the ‘one-to-ten rule’ (10 observations per variable). See Harrell *et al.* (1984); Harrell *et al.* (1996); Peduzzi *et al.* (1996); Laupacis *et al.* (1997) and Steyerberg *et al.* (2000) for more on this rule. Ultimately this helps to avoid overfitting. A k-fold cross validation would have split the data into two parts at least, thus creating unreliable overfitted models with each iteration.

Altogether, with 5 input variables and 50 observations, models could be trained with 10 observations left over for validation. Running several model iterations with random train-validate splits and assessing consistency, an idea of model accuracy was gotten.

Linear regression model

A linear regression model in the context of machine learning is one which mathematically patterns the association between one or more numerical predictors and a continuous response variable, for the purpose of predicting the response variable to a reasonable degree of accuracy, when applied to a set of non-modelled covariates.

A summary equation of the multiple linear regression model is:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \ \ i = 1, \ \ldots, n, \quad (1)$$

where: y_i is the *i*th response; β_k is the *k*th coefficient, where β_0 is the constant term in the model; X_{ij} is the *i*th observation on the *j*th predictor variable, $j = 1, \ldots, p$; ε_i is the *i*th noise expression, referring to random error.

All this operating under the assumptions that:

- The noise expressions, ε_i, are uncorrelated.
- The noise expressions, ε_i, have independent and similarly normal distributions with mean zero and constant variance, σ_2.
- The responses $y_i$are not correlated.

Least squares linear regression and robust linear regression were both explored.

In least squares regression, the coefficients are approximated to minimize the mean squared divergence between the predicted and actual response.

Equation for least squares linear regression;

$$\widehat{y_i} = \sum_{k=0}^{n} b_k f_k \left(X_{i1}, X_{i2}, \ \ \ldots, X_{ip}\right), \quad i = 1, \ \ldots, \ n, \quad (2)$$

where: $\widehat{y_i}$ is the response and b_k is the fitted coefficient.

The robust linear regression method is less affected by outliers than least squares regression and functions by assigning a weight to each data point using a technique called iteratively reweighted least squares (Barreto and Burrus 1994*a*, *b*; Burrus *et al.* 1994; Burrus 1998*a*, *b*). In the prime iteration of this process, every data point is equally weighted, and coefficients are approximated using least squares. In consequent iterations, weights are recalculated such that points that highly deviate from model predictions in prior iterations are assigned lower weighting. Model coefficients are then recalculated, applying weighted least squares. This workflow is repeated until resulting coefficients coincide around a prescribed tolerance.

For a detailed examination of the modalities and technicalities of linear regression modelling see Seber (1977); Neter *et al.* (1996); Bingham and Fry (2010) and Chatterjee and Simonoff (2012).

An example of a similar work employing multiple regression modelling for reservoir performance prediction is Oladeinde *et al.* (2015). In that instance a multiple linear regression model was created to forecast total production volume based on six predictors including gas–oil ratio, number of wells and a few well performance indices. The project was limited in scope but appeared to yield positive results.

Support vector regression (SVR) model

Support vector machine (SVM) analysis also referred to as support vector networks is a tool in supervised machine learning relevant to both classification and regression exercises (Gunn 1998). SVM typically refers to the use of support vectors for classification while SVR refers to regression specific support vectors (as used in this case). SVR here, as put forward by Vladimir Vapnik (see Vapnik 1995), uses an epsilon (ε)-insensitive loss function. ε referring to the distance of data points from the hyperplane, which in SVM would be the hyperplane of separation between groups of classes but in SVR operates as the line of prediction (or more precisely the midpoint of the margin of prediction). The best hyper-plane is the one with the greatest margin of prediction (boundary slab) between classes, which may not necessarily create a perfect distinction between classes but separates a substantial amount of the points; presenting what could be termed a soft margin.

In this form of regression modelling the raw data is mapped onto a higher dimensional space based on the chosen Kernel Function where the projected points closest to the boundary plane are termed support vectors. It can be described as non-linear mapping of projected input variables to create a linear predictive function (See Fig. 2 for illustration).

The linear support vector function would be mathematically expressed as:

$$f(x) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*)(x'_n x) + b \tag{3}$$

where: α_n and α_n^* are Lagrange multipliers; x terms represent support vectors; b is the bias term.

Apart from the Kernel Function, another key optimizable parameter in SVR is the box constraint. This parameter controls the strictness of datapoint classification, and the penalization imposed on misclassification; such that the higher the box constraint, the higher the cost of misclassification – leading to the designation of fewer support vectors and stricter data separation. The Kernel scale mode can also be optimized when a radial basis function kernel is applied. In this instance only the Linear and Cubic kernel functions are applied and so the kernel scale is not relevant.

Various workers have applied SVR in similar and adjacent contexts including Saffarzadeh and Shadizadeh (2012); Al-Anazi and Gates (2010); Gholami and Moradzadeh (2011) and Gholami *et al.* (2012) where it was applied to reservoir quality predictions and Zhong *et al.* (2010) where it was applied to predict production in high water cut fields.

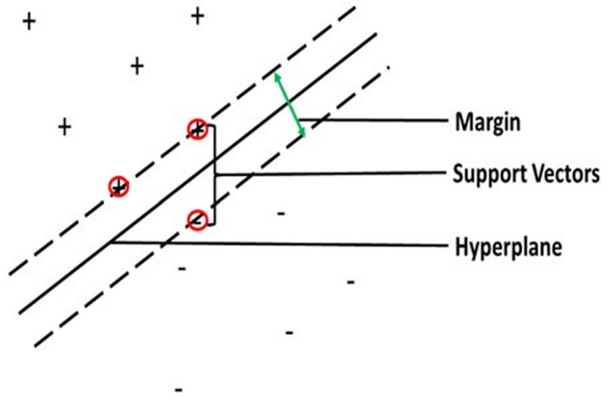

Fig. 2. Illustration of SVM classification method showing Support Vectors, data points from different classes (indicated by + or −), boundary slab (margin) indicated by dashed lines and hyperplane represented by solid line through the middle.

For a thorough exploration of the operating concepts for SVM see Steinwart and Christmann (2008).

Boosted regression tree model

Boosted regression tree modelling is a supervised learning technique that aggregates several models into a single predictive algorithm. This method specifically amalgamates the advantages of two processes. Namely, regression trees which relate a dependent (response) variable to its independent (control) variables by iterative twofold splits (see Fig. 3); and boosting which is an adaptive method for merging several uncomplicated models into one with more complexity, to improve forecasting ability (Elith *et al.* 2008).

Boosting is a stepwise process that seeks to minimize the loss function by including, at each tier, a new tree that best mitigates deviance.

Key user definable parameters for Boosted Tree Regression include:

- The Minimum Leaf Size: which equates to the complexity of each individual tree, with smaller leaf sizes more prone to recording noise in the data
- The Number of Trees: which equates to the number of learners to be aggregated
- The Learning Rate (Shrinkage parameter): a value between 0 and 1 which refers to the contribution of each tree to the model (the rate at which the model learns). Thus, the smaller the Learning rate the greater the number of iterations required (Hastie *et al.* 2009).

Elith *et al.* (2008) and Hastie *et al.* (2009) provide in-depth explanations on this technique. In an adjacent context to this project, Subasi *et al.* (2020) discuss the application of boosted trees to reservoir quality prediction.

Results

As outlined above, 5 different models were implemented (linear regression, robust linear regression, linear kernel support vector regression, cubic kernel support vector regression and boosted trees regression) using the aforementioned software. The outputs for presentation include graphs showing validation/test results of predicted responses v. actual (true) responses. These outputs are the result of random train-test splits of the data and are consistent regardless of split, with only slight variations observed across multiple iterations (5 or 6).

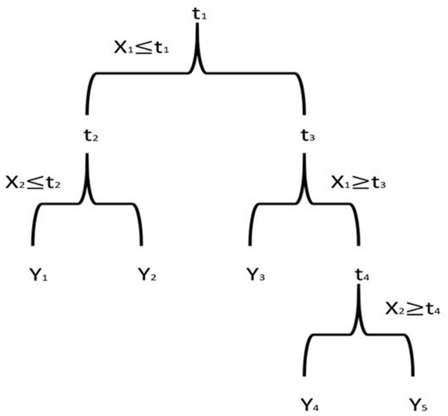

Fig. 3. A diagrammatic expression of a decision tree with a response Y, and independent variables X_1 and X_2, with splits t_1 to t_4 (after Elith *et al.* 2008).

Tables of model performance indicators are also shown below (Table 5)<CE: Please check table citations are not in sequential order>. These include:

- R-squared; a measure of the level of variation in the response variable explicable by the predictor variables. This percentage value is meant to indicate the predictive ability of the model. There is no universally agreed acceptable R-squared, but an R-squared value over 50% would be deemed acceptable (Bunge and Judson 2005).
- Mean Square Error (MSE) and Root Mean Square Error (RMSE): MSE (also known as the Mean Squared Deviation) is the average of the squares of the differences between the predicted and true responses. The RMSE is the square root of the MSE. These values are unit-less and provide a measure of the model accuracy. Being an average, this value is sensitive to outliers. To constrain RMSE values, both response variables of maximum field production rate and recovery factor were standardized based on maximum value to rescale outputs between 0 and 1. For RF all values were divided by 100, while values were divided by 300 for maximum field production rate. Thus, RMSE values are all <1.
- Mean Absolute Error (MAE): this is the average of the differences between predicted and true responses.

No universal cut-offs are recommended for error readings (RMSE, MSE and MAE). These values simply give a measure of how much predicted responses differ from actual responses on average.

Discussion

Values shown in Table 5 illustrate that there are readily observable differences in the performance of the models for the two outcome parameters.

For MFPR we observe that the least squares linear multiple regression model performs quite poorly. However, with a robust (iteratively reweighted least squares) multiple linear regression there is a spike in all measures of performance for predictability and accuracy, with over 80% R-squared and very low values of mean squared (MSE and RMSE) and absolute errors (see Table 5). A comparison of the predicted response v. true response graphs for both models (Fig. 4a, b respectively) illustrates the difference in predictive performance. The linear SVR also displays good performance (as reflected in Fig. 4c) with over 70% R-squared and RMSE less than 0.1 (Table 5). The cubic kernel SVR on the other hand performed poorly as a model for MFPR with skewed predicted response v. true response alignment (Fig. 4d). For the boosted regression model with minimum leaf size of 8 and 30 trees, model response of 65% R-squared and RMSE less than 0.1 was returned with visibly good prediction output (Fig. 4e).

For recovery factors we see that the cubic kernel SVR shows good R-squared of 65% and RMSE just under 0.1 and good correlation for predicted response v. true response (Fig. 5d). For other models, prediction of recovery factor was not as good as the Cubic SVR; RMSE values were over 0.1 and predicted response v. true response outputs (Fig. 5a–c and e) displayed more scatter.

Overall, it would appear MFPR is more easily predictable across a wider range of models than RF given the same predictor variables.

A closely similar experiment was discussed in Aliyuda *et al.* (2020). There, several predictive models were run on hydrocarbon (oil and gas) field data from the Norwegian continental shelf. Variables used in that study were mostly similar to the wider selection from this paper. However, relative to this study, no feature selection was applied there and hence 30 variables were processed through each model as compared to the 5 variables selected here based on PCA and subset testing. To compare results, three parameters were used as outcome variables in that paper, two of which match the two used here; specifically recovery factor and maximum field rates. Model performance metrics in that paper were also R-squared, RMSE and MSE.

In that study it was observed that support vector regression produced the best results for both recovery factor and maximum field rate (with maximum field rate also having a better R-squared than recovery factor) as was observed here. That paper did not explicitly state which kernel function (whether linear, cubic, quadratic, etc.) was used in the SVRs for the testing of each outcome variable.

Results from this study show that with the data used for training the model, recovery factor could be predicted up to as high as 65% R-squared using a cubic kernel support vector regression method with very low absolute error equating to within single digit percentages of recovery factor. Poorest performance in recovery factor prediction is the linear regression model at −22% R-squared. Maximum field production rate could also be predicted with a high level of certainty with models producing up to 85% R-squared and with low absolute errors (less than single standardized unit) in models with good performance (above 50% R-squared). Observing the results, a recommendation is made for the use of support vector regression in reservoir/field performance prediction, with tuning of kernel functions depending on the outcome variable being predicted.

As to why the different algorithms produce different results, simply put, it is like applying a mathematical function or constant to the exact same data. A basic example of this concept would be having an addition function (‘+’) and a multiplication function (‘x’) and two numbers e.g. 4 and 7. Applying the multiplication function to the two numbers would equate to 28 (4 × 7), while applying the addition function would equate to 11 (4 + 7). Same data sample,

Table 5. *Table of results showing resulting measures of accuracy and predictability for the supervised learning models of the two response variables of MFPR and RF*

		Model				
Response variable	Metric	Linear	Robust linear	Linear SVR	Cubic SVR	Boosted regression
Maximum Field Production Rate	**R-Squared (%)**	−29	85	76	−42	65
	RMSE	0.1587	0.0543	0.0692	0.1670	0.0824
	MSE	0.0252	0.0030	0.0048	0.0279	0.0068
	MAE	0.1228	0.0444	0.0587	0.1401	0.0584
Recovery Factor	**R-Squared (%)**	−22	−23	−16	65	27
	RMSE	0.1708	0.1714	0.1663	0.0916	0.1321
	MSE	0.0292	0.0294	0.0276	0.0084	0.0175
	MAE	0.1412	0.1420	0.1380	0.0734	0.1078

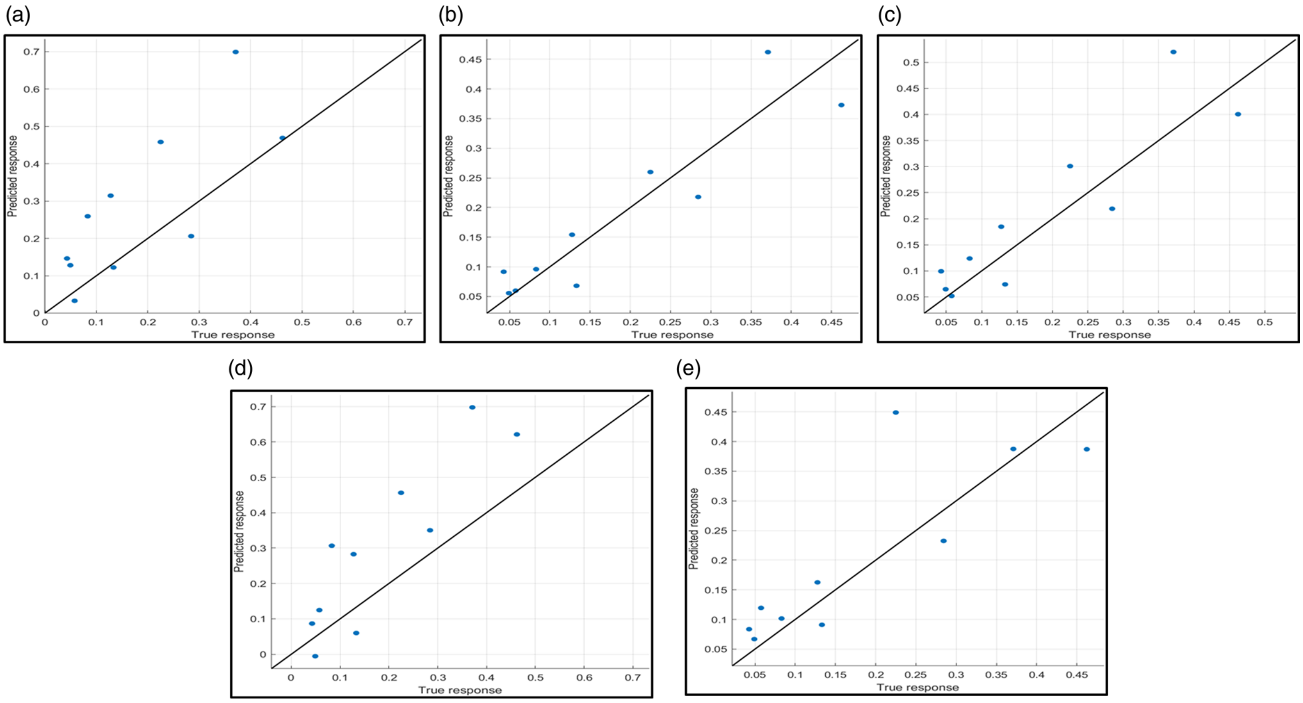

Fig. 4. These figures show predicted v. response graphs for machine learning models of maximum field production rate: (**a**) Linear regression model; (**b**) Robust linear regression model; (**c**) Linear support vector regression model; (**d**) Cubic support vector regression model; (**e**) Boosted tree regression model.

different processes applied to create different results. Even so on a more complex level of algorithmic processes the model equations take the same basic data and functionally wrangle them differently.

In this specific instance the key difference in the way SVR processes the data and other regression methods do, is that SVR creates a broad range of fit (as captured in the boundary slab illustrated in Fig 4.2) in its predictive process resulting in what some refer to as a 'low bias and high variance' model; while linear regression methods create a single line of best fit through points thus presenting a 'high bias and low variance' prediction. Where the boosted regression thrives is that it aggregates multiple regressive processes broadening the otherwise high bias low variance situation presented by a singular linear regression.

The implication of this work is that with an abundance of legacy data floating around in E&P industry and academia not only can insights into the production process be acquired but with a small number of variables production performance can reliably be predicted.

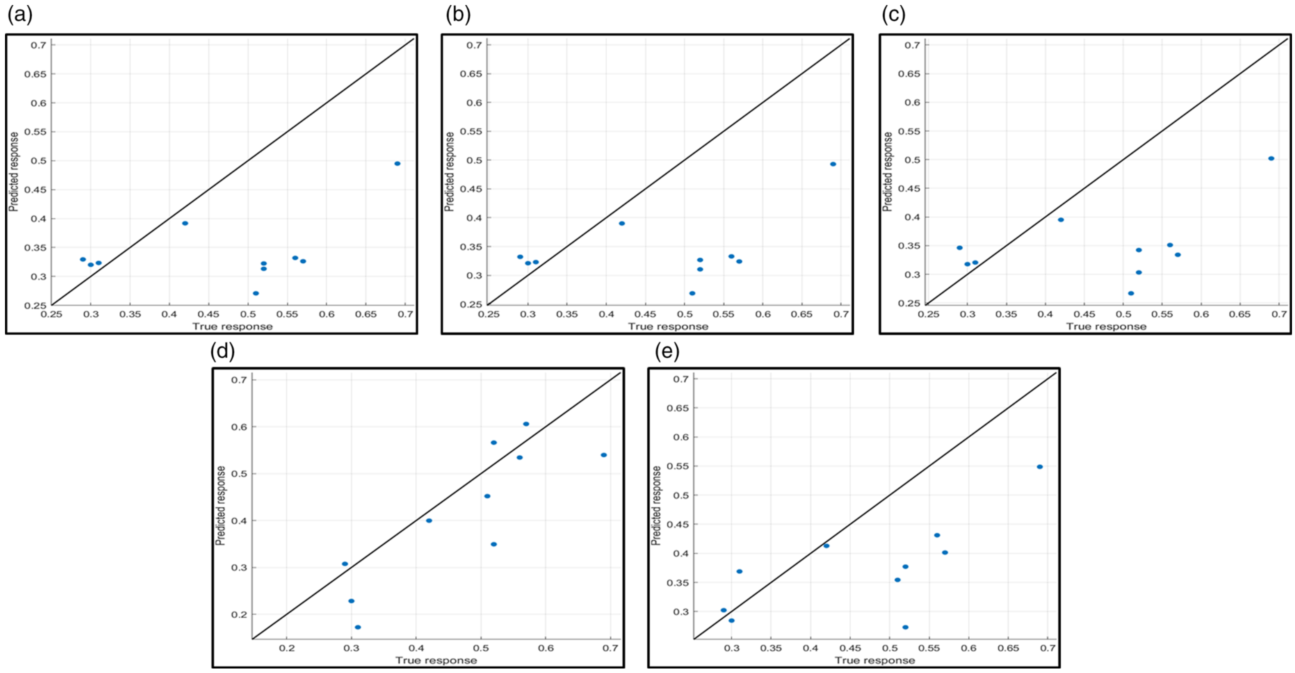

Fig. 5. These figures show predicted v. true response graphs for machine learning models of recovery factor: (**a**) Linear regression model; (**b**) Robust linear regression model; (**c**) Linear support vector regression model; (**d**) Cubic support vector regression model; (**e**) Boosted tree regression model.

Another paper dealing with prediction of reservoir performance using other machine leaning methodologies is Panja *et al.* (2018) where two machine learning models (least squares support vector machine and artificial neural networks) were tested in the prediction of oil recovery against a curve fitting model using simulated data generated for 8 variables (permeability, initial dissolved GOR, rock compressibility, gas relative permeability, slope of GOR, initial pressure, flowing bottom hole pressure, and hydraulic fracture spacing) with 114 observations to train and 30 observations to test. Notably, the control variables somewhat intersect with the selection used in this paper in terms of permeability and GOR. The data was used to predict recovery factor (also similar to this study) as well as produced gas–oil ratio. The results showed, in agreement with this study, that SVM is quite accurate for predicting recovery factor.

Many other studies exist showing similarities and differences in methodologies and variable selection to those used here to achieve high coefficients of determination (R-squared), including Belazreg *et al.* (2019) and Belazreg *et al.* (2021) where predictive models for recovery factor were developed based on regression and group method of data handling (GMDH) with positive results (R-squared as high as 72%). An exhaustive rundown of all these studies and their intricacies would not be feasible. The important thing to note is that these various methods of machine learning prove to be successful in these case studies, using few variables to predict various measures of reservoir performance.

A few other papers dealing with the subject include Mohammadi *et al.* (2014), Srivastava *et al.* (2016) and Daribayev *et al.* (2020).

In summary, numerous papers on machine learning application for performance prediction exist (as referenced in the literature review of this paper), each dealing with unique case studies and employing a variety of artificial intelligence methodologies. This paper is an addition to that ever-expanding library, with its detailing of the application of real-world data for prediction algorithm utilization.

Conclusion

From observations of results, we see that statistics based predictive models can be used to provide accurate reservoir performance forecasting. It is also apparent that depending on the outcome being predicted (using the exact same predictor variables), the model being applied might require adjustment of tuning parameters or the use of a different model altogether.

Comparing results for the two responses (RF and MFPR) from this study as well as previously published studies it would appear that SVR is a good modelling technique for reservoir performance prediction overall. For different response variables, a change in kernel function (linear, cubic, gaussian, etc.) should produce high R-squared and low error.

Future work would involve broadening the scope of data in terms of number of observations, training reservoir data from different hydrocarbon producing regions and assessing other response variables to decipher empirically appropriate algorithms.

Acknowledgements Classification schema for Gross Depositional Environment is applied from the SAFARI Database Project (safaridb.com).

Author contributions **UO**: data curation (lead), formal analysis (lead), funding acquisition (lead), methodology (equal), validation (lead), visualization (equal), writing – original draft (lead); **JH**: conceptualization (lead), methodology (equal), supervision (lead), visualization (equal), writing – review & editing (lead)

Funding This work was funded by the Petroleum Technology Development Fund (PTDF/ED/OSS/PHD/OU/1188/17).

Competing interests 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.

Data availability Data for this research were acquired from a variety of public sources. The curated database along with a complete list of references is available from the University of Aberdeen Library Cataloguing Service (cataloguing@abdn.ac.uk), and on request from the corresponding Author.

References

Ahmadi, M., Soleimani, R., Lee, M., Kashiwao, T. and Bahadori, A. 2015. Determination of oil well production performance using Artificial Neural Network (ANN) linked to the particle swarm optimization (PSO) Tool. *Petroleum*, **1**, 118–132, https://doi.org/10.1016/j.petlm.2015.06.004

Ahmed, T. and Mckinney, P. 2005. Predicting oil reservoir performance. *Advanced Reservoir Engineering*, **1**, 327–363, https://doi.org/10.1016/B978-075067733-2/50007-1

Ahmed, T. and Meehan, D. 2012. Predicting oil reservoir performance. *Advanced Reservoir Management and Engineering*, **2**, 485–539, https://doi.org/10.1016/B978-0-12-385548-0.00005-1

Al-Anazi, A.F. and Gates, I.D. 2010. Support vector regression for porosity prediction in a heterogeneous reservoir: a comparative study. *Computers and Geosciences*, **36**, 1494–1503, https://doi.org/10.1016/j.cageo.2010.03.022

Al-Fattah, S. and Startzman, R. 2001. Predicting Natural Gas Production using Artificial Neural Network. SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas.

Aliyuda, K., Howell, J. and Humphrey, E. 2020. Impact of geological variables in controlling oil-reservoir performance: an insight from a machine-learning technique. *SPE Reservoir Evaluation & Engineering*, **23**, 1314–1327, https://doi.org/10.2118/201196-PA

Al-Jifri, M., Al-Attar, H. and Boukadi, F. 2021. New proxy models for predicting oil recovery factor in waterflooded heterogeneous reservoirs. *Journal of Petroleum Exploration and Production Technology*, **11**, 1443–1459, https://doi.org/10.1007/s13202-021-01095-4

Amirian, E., Leung, J., Zanon, S. and Dzurman, P. 2013. Data-driven modeling approach for recovery performance prediction in SAGD operations. SPE Heavy Oil Conference-Canada, Calgary, Alberta, Canada.

Ärlebrand, A., Augustsson, C., Escalona, A., Grundvåg, S. and Marín, D. 2021. Provenance, depositional setting and diagenesis as keys to reservoir quality of the lower cretaceous in the SW Barents Sea. *Marine and Petroleum Geology*, **132**, 105217, https://doi.org/10.1016/j.marpetgeo.2021.105217

Armitage, P., Worden, R., Faulkner, D., Aplin, A., Butcher, A. and Iliffe, J. 2010. Diagenetic and sedimentary controls on porosity in lower carboniferous fine-grained lithologies, Krechba field, Algeria: a petrological study of a caprock to a carbon capture site. *Marine and Petroleum Geology*, **27**, 1395–1410, https://doi.org/10.1016/j.marpetgeo.2010.03.018

Barreto, J.A. and Burrus, C.S. 1994*a*. Complex Approximation using Iterative Reweighted Least Squares for Fir Digital Filters. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages III: 545–548, IEEE ICASSP-94, April 19–22, Adelaide, Australia.

Barreto, J.A. and Burrus, C.S. 1994*b*. Iterative Reweighted Least Squares and the Design of Two-Dimensional Fir Digital Filters. Proceedings of the IEEE International Conference on Image Processing, volume I, pages I: 775–779, IEEE ICIP-94, November 13–16, Austin, Texas.

Belazreg, L., Mahmood, S. and Aulia, A. 2019. Novel approach for predicting water alternating gas injection recovery factor. *Journal of Petroleum Exploration and Production Technology*, **9**, 2893–2910, https://doi.org/10.1007/s13202-019-0673-2

Belazreg, L., Mahmood, S. and Aulia, A. 2021. Fast and cost-effective mathematical models for hydrocarbon-immiscible water alternating gas incremental recovery factor prediction. *ACS Omega*, **6**, 17492–17500, https://doi.org/10.1021/acsomega.1c01901

Bhattacharyya, S. and Vyas, A. 2022*a*. Application of machine learning in predicting oil rate. *Scientific Reports*, **12**, https://doi.org/10.1038/s41598-022-20401-6

Bhattacharyya, S. and Vyas, A. 2022*b*. Machine learning based rate decline prediction in unconventional reservoirs. *Upstream Oil and Gas Technology*, **8**, 100064, https://doi.org/10.1016/j.upstre.2022.100064

Bhattacharya, S., Ghahfarokhi, P., Carr, T. and Pantaleone, S. 2019. Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: a case study from the Marcellus Shale, North America. *Journal of Petroleum Science and Engineering*, **176**, 702–715, https://doi.org/10.1016/j.petrol.2019.01.013

Bingham, N.H. and Fry, J.M. 2010. *Regression Linear Models in Statistics*, 1st edn. Springer London, London.

Bunge, J. and Judson, D. 2005. Data mining. *Encyclopedia of Social Measurement*, **1**, 617–624, https://doi.org/10.1016/B0-12-369398-5/00159-6

Burrus, C.S. 1998*a*. Constrained Least Squares Design of Fir Filters Using Iterative Reweighted Least Squares. Proceedings of EUSIPCO-98, September 8–11, Rhodes, Greece. 281–282.

Burrus, C.S. 1998*b*. Convergence of constrained optimal design of fir filters using iterative reweighted least squares. Proceedings of the IEEE DSP Workshop – 1998, August 9–12, Bryce Canyon, Utah. paper #113.

Burrus, C.S., Barreto, J.A. and Selesnick, I.W. 1994. Iterative reweighted least squares design of fir filters. *IEEE Transactions on Signal Processing*, **42**, 29262936, https://doi.org/10.1109/78.330353

Busahmin, B. and Maini, B. 2010. Effect of solution-gas-oil-ratio on performance of solution gas drive in foamy heavy oil systems. Canadian Unconventional Resources and International Petroleum Conference.

Chatterjee, S. and Simonoff, J.S. 2012. *Handbook of Regression Analysis*. John Wiley & Sons, Incorporated.

Chithra Chakra, N., Song, K., Gupta, M. and Saraf, D. 2013*a*. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs). *Journal of Petroleum Science and Engineering*, **106**, 18–33, https://doi.org/10.1016/j.petrol.2013.03.004

Chithra Chakra, N., Song, K., Saraf, D.N. and Gupta, M. 2013*b*. Production forecasting of petroleum reservoir applying higher-order neural networks (HONN) with limited reservoir data. *International Journal of Computer Applications*, **72**, 23–35, https://doi.org/10.5120/12466-8834

Choubineh, A., Ghorbani, H., Wood, D., Robab Moosavi, S., Khalafi, E. and Sadatshojaei, E. 2017. Improved predictions of wellhead choke liquid critical-flow rates: modelling based on hybrid neural network training learning based optimization. *Fuel*, **207**, 547–560, https://doi.org/10.1016/j.fuel.2017.06.131

Dake, L.P. 1994. *The Practice of Reservoir Engineering: Developments in Petroleum Science*. Elsevier, Amsterdam, **36**.

Daribayev, B., Akhmed-Zaki, D., Imankulov, T., Nurakhov, Y. and Kenzhebek, Y. 2020. Using machine learning methods for oil recovery prediction. European Association of Geoscientists & Engineers. *ECMOR XVII*, **2020**, 1–13, https://doi.org/10.3997/2214-4609.202035233

Elith, J., Leathwick, J. and Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813, https://doi.org/10.1111/j.1365-2656.2008.01390.x

Ertekin, T. and Sun, Q. 2019. Artificial intelligence applications in reservoir engineering: a status check. *Energies*, **12**, 2897, https://doi.org/10.3390/en12152897

Ertekin, T., Abou-Kassem, J.H. and King, G.R. 2001. *Basic Applied Reservoir Simulation*. Society of Petroleum Engineers, Richardson, TX, 406.

Ghahfarokhi, P., Carr, T., Bhattacharya, S., Elliott, J., Shahkarami, A. and Martin, K. 2018. A fiber-optic assisted multilayer perceptron reservoir production modeling: a machine learning approach in prediction of gas production from the Marcellus Shale. *Proceedings of the 6th Unconventional Resources Technology Conference held in Houston, Texas*, USA, 23–25 July 2018.

Gholami, R. and Moradzadeh, A. 2011. Support vector regression for prediction of gas reservoirs permeability. *Journal of Mining and Environment*, **2**, 41–52, https://doi.org/10.22044/jme.2012.18

Gholami, R., Shahraki, A. and Jamali P.M. 2012. Prediction of hydrocarbon reservoirs permeability using support vector machine. *Mathematical Problems in Engineering*, **2012**, 1–18, https://doi.org/10.1155/2012/670723

Ghorbani, H., Wood, D., Moghadasi, J., Choubineh, A., Abdizadeh, P. and Mohamadian, N. 2019. Predicting liquid flow-rate performance through wellhead chokes with genetic and solver optimizers: an oil field case study. *Journal of Petroleum Exploration and Production Technology*, **9**, 1355–1373, https://doi.org/10.1007/s13202-018-0532-6

Gunn, S.R. 1998. *Support Vector Machines for Classification and Regression*. ISIS Technical Report 14.

Gunter, G., Finneran, J., Hartmann, D. and Miller, J. 1997. Early determination of reservoir flow units using an integrated petrophysical method. *SPE Annual Technical Conference and Exhibition*, San Antonio, Texas, USA, October 1997.

Gurbanov, R., Musayeva, S., Gurbanov, R. and Ahmedov, Z. 2016. Advanced well spacing system application in the development of oil and gas fields. *Procedia Computer Science*, **102**, 446–452, https://doi.org/10.1016/j.procs.2016.09.425

Han, D. and Kwon, S. 2021. Application of machine learning method of data-driven deep learning model to predict well production rate in the Shale Gas Reservoirs. *Energies*, **14**, 3629, https://doi.org/10.3390/en14123629

Harrell, F.E., Jr., Lee, K.L., Califf, R.M., Pryor, D.B. and Rosati, R.A. 1984. Regression modelling strategies for improved prognostic prediction. *Stat Med*, **3**, 143–152, https://doi.org/10.1002/sim.4780030207

Harrell, F.E., Jr., Lee, K.L. and Mark, D.B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, **15**, 361–387, https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4 < 361::AID-SIM168>3.0.CO;2-4

Hartmann, D.J. and Beaumont, E.A. 1999. Chapter 9, predicting reservoir system quality and performance. *In*: Beaumont, E.A. and Foster, N.H. (eds) *Exploring for Oil and Gas Traps*. Treatise of Petroleum Geology, Handbook of Petroleum Geology, 9-1–9-154.

Hastie, T., Tibshirani, R. and Friedman, J.H. 2009. *The Elements of Statistical Learning. Springer Series in Statistics*, 2nd edn. Springer, New York.

Ingles, M. and Anadon, P. 1991. Relationship of clay minerals to depositional environment in the Non-Marine Eocene Pontils Group, Se Ebro Basin (Spain). *SEPM Journal of Sedimentary Research*, **61**, 926–939.

Jolliffe, I. 2002. *Principal Component Analysis*, 2nd edn. Springer-Verlag, New York, NY.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2 (IJCAI'95) Montreal, Quebec, Canada, 20–25 August 1995. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137–1143.

Lai, J., Wang, G., Ran, Y. and Zhou, Z. 2015. Predictive distribution of high-quality reservoirs of tight gas sandstones by linking diagenesis to depositional facies: evidence from Xu-2 sandstones in the Penglai Area of the Central Sichuan Basin, China. *Journal of Natural Gas Science and Engineering*, **23**, 97–111, https://doi.org/10.1016/j.jngse.2015.01.026

Larue, D.K. and Friedmann, F. 2005. The controversy concerning stratigraphic architecture of channelized reservoirs and recovery by Waterflooding. *Petroleum Geoscience*, **11**, 131–146, https://doi.org/10.1144/1354-079304-626

Laupacis, A., Sekar, N. and Stiell, I.G. 1997. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA*, **277**, 488–494, https://doi.org/10.1001/jama.1997.03540300056034

Levell, B. 2021. Petroleum geoscience. *Encyclopedia of Geology*, **5**, 762–782, https://doi.org/10.1016/B978-0-08-102908-4.00073-4

Lever, J., Krzywinski, M. and Altman, N. 2017. Principal component analysis. *Nature Methods*, **14**, 641–642, https://doi.org/10.1038/nmeth.4346

Li, X., Chan, C. and Nguyen, H. 2013. Application of the neural decision tree approach for prediction of petroleum production. *Journal of Petroleum Science and Engineering*, **104**, 11–16, https://doi.org/10.1016/j.petrol.2013.03.018

Liu, W., Liu, W. and Gu, J. 2020. Forecasting oil production using ensemble empirical model decomposition based long short-term memory neural network. *Journal of Petroleum Science and Engineering*, **189**, 107013, https://doi.org/10.1016/j.petrol.2020.107013

Lorenz, J.C., Sattler, A.R. and Stein, C.L. 1989. The effects of depositional environment on petrophysical properties of mesaverde reservoirs, Northwestern Colorado. *Proceedings of the 64th Society of Petroleum Engineers Annual Technical Conference and Exhibition*, San Antonio, Texas, USA, 119–132, https://doi.org/10.2118/SPE-19583-MS

Manzir, M.P., Beka, F.T. and Kadana, R.I. 2015. Predicting reservoir performance changes with time. *International Journal for Research in Emerging Science and Technology*, **2**, 85–94, https://ijrest.net/downloads/volume-2/issue-9/pid-ijrest-29201514.pdf

Meehan, D. 2011. Using analog reservoir performance to understand EOR opportunities in type I fractured reservoirs. *SPE Enhanced Oil Recovery Conference*, Kuala Lumpur, Malaysia, July 2011, https://doi.org/10.2118/144177-MS.

Mirzaei-Paiaman, A. and Salavati, S. 2012. The application of artificial neural networks for the prediction of oil production flow rate. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, **34**, 1834–1843, https://doi.org/10.1080/15567036.2010.492386

Mitchell, T. 1997. *Machine Learning*. McGraw Hill, **2**. ISBN 978-0-07-042807-2.

Mode, A.W., Anyiam, O.A. and John, S.I. 2017. Depositional environment and reservoir quality assessment of the 'Bruks Field,' Niger Delta. *Journal of Petroleum Exploration and Production Technology*, **7**, 991–1002, https://doi.org/10.1007/s13202-017-0346-y

Mohammadi, M., Kouhi, M. and Mohebbi, A. 2014. Prediction of oil recovery factor in CO2 injection process. *Petroleum Science and Technology*, **32**, 2093–2101, https://doi.org/10.1080/10916466.2012.743563

Mustafiz, S. and Islam, M.R. 2008. State-of-the-art petroleum reservoir simulation. *Petroleum Science and Technology*, **26**, 1303–1329, https://doi.org/10.1080/10916460701834036

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. 1996. *Applied Linear Statistical Models*. IRWIN, The McGraw-Hill Companies, Inc.

Niu, W., Lu, J. and Sun, Y. 2021. A production prediction method for shale gas wells based on multiple regression. *Energies*, **14**, 1461, https://doi.org/10.3390/en14051461

Okotie, S. and Ikporo, B. 2018. Reservoir performance prediction. *Reservoir Engineering*, Springer, Cham, **1**, 365–410, https://doi.org/10.1007/978-3-030-02393-5_11.

Oladeinde, M., Ohwo, A. and Oladeinde, C. 2015. A mathematical model for predicting output in an oilfield in the Niger Delta Area of Nigeria. *Nigerian Journal of Technology*, **34**, 768, https://doi.org/10.4314/njt.v34i4.14

Pandey, R., Dahiya, A. and Mandal, A. 2020. Identifying applications of machine learning and data analytics based approaches for optimization of upstream petroleum operations. *Energy Technology*, **9**, 2000749, https://doi.org/10.1002/ente.202000749

Panja, P., Velasco, R., Pathak, M. and Deo, M. 2018. Application of artificial intelligence to forecast hydrocarbon production from Shales. *Petroleum*, **4**, 75–89, https://doi.org/10.1016/j.petlm.2017.11.003

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49**, 1373–1379, https://doi.org/10.1016/S0895-4356(96)00236-3

Reinson, G.E. 1991. Depositional Facies Control of Reservoir Heterogeneity and Performance, Cretaceous Sandstone Reservoirs, South-Central Alberta Basin. Geological Survey of Canada, Calgary, Alberta, Canada. AAPG Search and

Discovery Article #91004 © 1991 AAPG Annual Convention Dallas, 7–10 April 1991, Texas.

Richards, M. and Bowman, M. 1998. Submarine fans and related depositional systems II: variability in reservoir architecture and wireline log character. *Marine and Petroleum Geology*, **15**, 821–839, https://doi.org/10.1016/S0264-8172(98)00042-7

Russell, S.J. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*, 3rd edn. Prentice Hall.

Saffarzadeh, S. and Shadizadeh, S. 2012. Reservoir rock permeability prediction using support vector regression in an Iranian oil field. *Journal of Geophysics and Engineering*, **9**, 336–344, https://doi.org/10.1088/1742-2132/9/3/336

Seber, G.A.F. 1977. *Linear Regression Analysis. Wiley Series in Probability and Mathematical Statistics*. John Wiley and Sons, Inc.

Shaw, T., Gieskes, J. and Jahnke, R. 1990. Early diagenesis in differing depositional environments: the response of transition metals in pore water. *Geochimica et Cosmochimica Acta*, **54**, 1233–1246, https://doi.org/10.1016/0016-7037(90)90149-F

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N. and Oza, H. 2021. Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research* **6**, 379–391, https://doi.org/10.1016/j.ptlrs.2021.05.009

Srivastava, P., Wu, X., Amirlatifi, A. and Devegowda, D. 2016. Recovery Factor Prediction for Deepwater Gulf of Mexico Oilfields by Integration of Dimensionless Numbers with Data Mining Techniques. SPE Intelligent Energy International Conference and Exhibition, Aberdeen, Scotland.

Steinwart, I. and Christmann, A. 2008. *Support Vector Machines*. Springer-Verlag, New York.

Steyerberg, E.W., Eijkemans, M.J., Harrell, F.E., Jr. and Habbema, J.D. 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*, **19**, 1059–1079, https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8 < 1059::AID-SIM412>3.0.CO;2-0

Subasi, A., El-Amin, M., Darwich, T. and Dossary, M. 2020. Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression. *Journal of Ambient Intelligence and Humanized Computing* **13**, 3555–3564, https://doi.org/10.1007/s12652-020-01986-0

Toevs, G., Morra, M., Winowiecki, L., Strawn, D., Polizzotto, M. and Fendorf, S. 2008. Depositional influences on porewater arsenic in sediments of a mining-contaminated freshwater lake. *Environmental Science & Technology*, **42**, 6823–6829, https://doi.org/10.1021/es800937t

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

Wang, A., Zhong, D. *et al.* 2018. Depositional and diagenetic controls on the reservoir quality of upper Triassic Chang-7 tight oil sandstones, Southwestern Ordos Basin, China. *Geosciences Journal*, **23**, 471–488, https://doi.org/10.1007/s12303-018-0042-z

Zhang, J., Qin, L. and Zhang, Z. 2008. Depositional facies, diagenesis and their impact on the reservoir quality of silurian sandstones from Tazhong Area in Central Tarim Basin, Western China. *Journal of Asian Earth Sciences*, **33**, 42–60, https://doi.org/10.1016/j.jseaes.2007.10.021

Zhong, Y., Zhao, L., Liu, Z., Xu, Y. and Li, R. 2010. Using a support vector machine method to predict the development indices of very high water cut oilfields. *Petroleum Science*, **7**, 379–384, https://doi.org/10.1007/s12182-010-0081-1