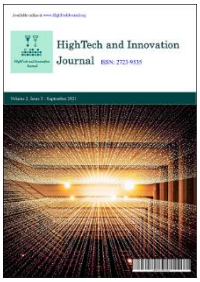
Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 2, No. 3, September, 2021



# A Data Mining Perspective on the Confluent Ions` Effect for Target Functionality

Babak Fazelabdolabadi [a*], Mostafa Montazeri [a], Peyman Pourafshary [b]

[a] *Center for Exploration and Production Studies and Research, Research Institute of Petroleum Industry (RIPI), Tehran, Iran.*

[b] *School of Mining and Geosciences, Nazarbayev University, 53 Kabanbay Batyr Avenue, Nur-Sultan City, Kazakhstan.*

Received 08 February 2021; Revised 11 May 2021; Accepted 24 June 2021; Published 01 September 2021

**Abstract**

The production of hydrocarbon resources at an oil field is concomitant with challenges with respect to the formation of scale inside the reservoir rock, intricately impairing its permeability and hindering the flow. Historically, the effect of ions has been attributed to the undergone phenomenon; nevertheless, there exists a great deal of ambiguity about its relative significance compared to other factors, or the effectiveness as per the ion type. The present work applies a data mining strategy to uncover the influence hierarchy of the parameters involved in driving the process within major rock categories—sandstone and carbonate—to regulate a target functionality. The functionalities considered revolve around maximizing oil recovery and minimizing permeability impairment/scale damage. A pool of experimental as well as field data was used for this purpose, accumulating the bulk of the available literature data. The methods used for data analysis in the present work included the Bayesian Network, Random Forest, Deep Neural Network, as well as Recursive Partitioning. The results indicate a rolling importance for different ion species, altering under each functionality, which is not ranked as the most influential parameter in either case. For the oil recovery target, our results quantify a distinction between the source of an ion of a single type in terms of its influencing rank in the process. This latter deduction is the first proposal of its kind, suggesting a new perspective for research. Moreover, the machine learning methodology was found to be capable of reliably capturing the data, as evidenced by the minimal errors in the bootstrapped results.

*Keywords:* Big Data; Machine Learning; Bayesian Networks; Random Forest; Formation Damage; Oil Recovery.

## 1. Introduction

As a long-lasting issue in petroleum production, the formation of scales continues to impede the flow and cause an economic burden on the upstream sector, which is estimated to be in excess of billions of dollars worldwide [1]. The formation of scales subsequently reduces the permeability of the formation [2], which adversely affects the recovery of the hydrocarbon resources. Given its importance, several studies have focused on understanding the effects of different parameters on the scale formation phenomenon and proposing relevant mechanisms. According to the literature, the scale formation at an oilfield is linked with operational parameters such as field type as well as laboratory parameters such as the concentration of selected ions, with the latter being tested both inside inhibitor-free and inhibitor-containing environments [3-6]. Nevertheless, the theories evolved over the deposition mechanism are non-overlapping and there exists a great deal of ambiguity about the influencing rank of the found parameters in the sequel—a question which this work attempts to address.

* Corresponding author: bkfazel@yahoo.com

http://dx.doi.org/10.28991/HIJ-2021-02-03-05

A plausible deduction of the actual interplay between parameters affecting the formation damage/oil recovery process can be merely derived by considering all the parameters involved simultaneously. This definition refers to a big-data framework, representing a favorably large sample size, to be subsequently applied to machine learning strategies. In practice, however, the available data in the literature adheres to a study conducted on a given functionality—maximizing the oil recovery while minimizing permeability impairment/scale damage. It is therefore logical to construct a specific database for each target functionality, for which the data is separately available.

The application of machine learning strategies has been widely practiced in oil and gas development. These attempts have covered aspects of enhanced oil recovery [7–14], fracture detection [15], development plan optimization [15, 16], dynamic production prediction [18–21] and asphaltene precipitation prediction [22]. Some studies have also focused on applying machine learning strategies to model permeability impairment due to mineral scale deposition [23–25] and predict the success of an inhibition scenario in the field [4]. The bulk of these works have adopted an Artificial Neural Network (ANN) technique for their analysis [26], albeit some hybrid strategies have also been tested. In essence, these hybrid methods were created by introducing an adjustment to the neuron weights (inside ANN settings) through meta-heuristics algorithms, such as the Imperialist Competitive Algorithm (ICA), Genetics algorithm (GA), particle swarm optimization (PSO), or both (HGAPSO). The modifications have reportedly yielded improved overall accuracy; nevertheless, some developed models bear limitations with respect to being trapped within local optima, making their predictions unreliable for a certain range of the data spectrum [23]. This creates a necessity for other machine learning techniques to be also evaluated for the same target.

A common feature of the recent machine-learning investigations on the reservoir mineral scale prediction [23-25] has been the adoption of a single data-bank as their model input, which reports experimental results on sandstone rocks. This brings about another limitation to the established results algorithm efficiency or parameter importance rank as being specific to the given rock type, or being tested otherwise. The present work contributes to the existing literature in this field in several ways testing new algorithms efficiency within sandstone and carbonate rocks, providing an in-depth view of parameter importance rank for a specific functionality. In this regards, the authors have also accumulated a data-set of experimental results on oil recovery from carbonate rocks from both the literature and our experiments with added parameters list, to include the source of ions, for further importance level classification deducted from data processing. A flowchart is presented in Figure 1 to illustrate our research methodology.

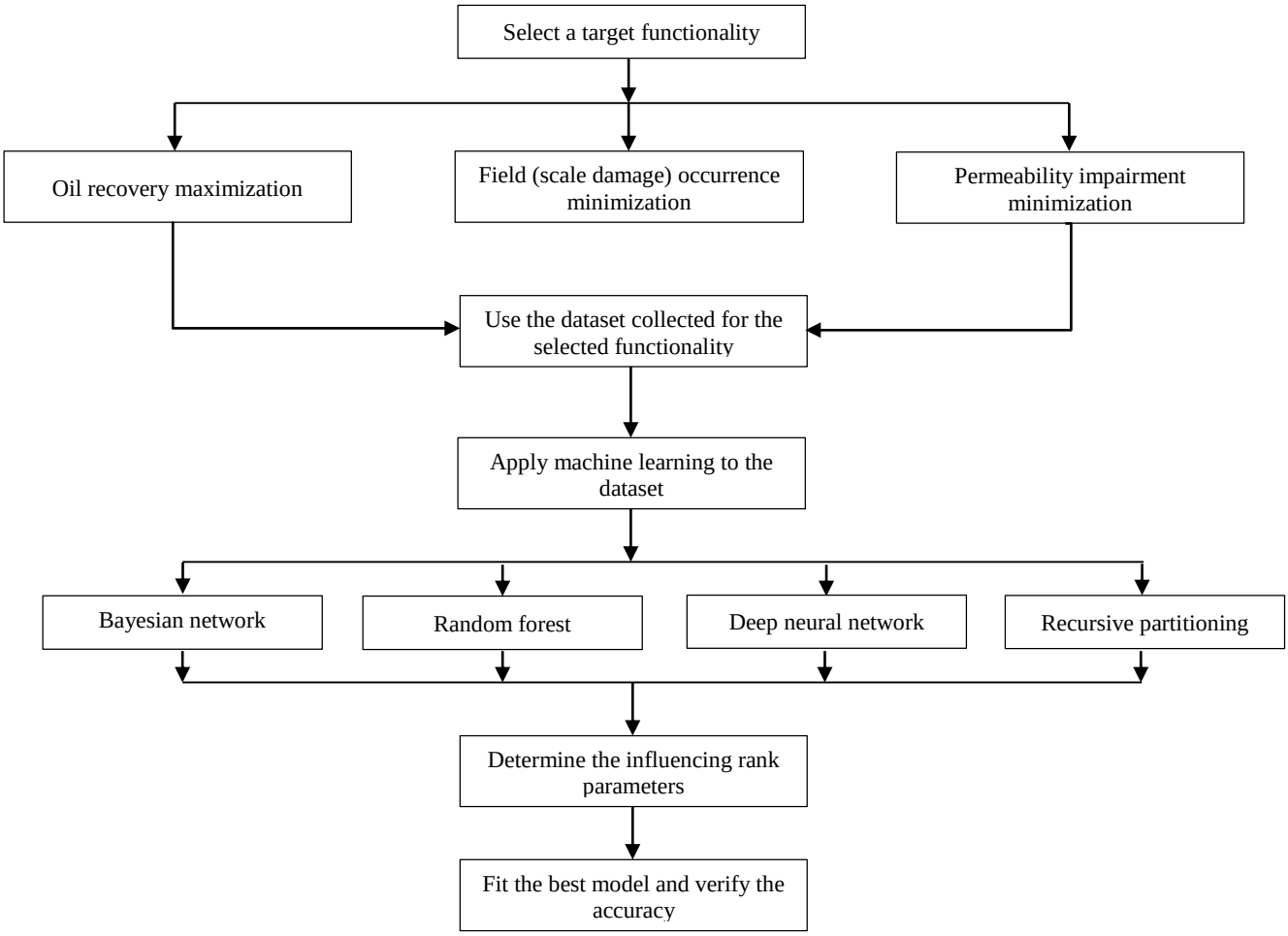



**Figure 1. The flowchart of the research methodology followed in the present work**

## 2. Data

The data used in the present work was obtained from the open literature as well as our own experiments [4, 25, 27-30]. As explained earlier, the data was collected so as to target three main functionalities minimizing permeability impairment (I), minimizing the possibility of scale damage in the field (II), maximizing oil recovery from matrix (III). As such, three distinct sets of data were acquired with essentially different parameters list. The list alters slightly under each category owing to their original recording scheme. In essence, the lists include parameters related to the fluid/matrix properties as well as the experimental/field conditions, under which the data was obtained. Tables 1 to 3 provide a description of the parameters considered under each functionality. The embedding rock type for the data in category (I) belongs to sandstone; whereas in the other two categories the data refers to the carbonate case.

**Table 1. The description of parameters used for target functionality (I) - permeability impairment**

| Name | Description (unit) | |
|---|---|---|
| x1 | pore volume | (-) |
| x2 | Rate of injection | (cc/min) |
| x3 | Temperature | (C) |
| x4 | Pressure difference along the core | (psi) |
| x5 | Initial permeability | (md) |
| x6 | $Ba_2+$ ion concentration in formation water | (ppm) |
| x7 | $Sr_2+$ ion concentration in formation water | (ppm) |
| x8 | $Ca_2+$ ion concentration in formation water | (ppm) |
| x9 | $SO_4(_2-)$ ion concentration in formation water | (ppm) |
| x10 | Final permeability, experimental | (md) |

**Table 2. The description of parameters used for target functionality (II) - $CaCO_3$ scale damage possibility**

| Item | Description (unit) | |
|---|---|---|
| 1 | Field | (-) |
| 2 | Na+ ion concentration in injecting fluid | (ppm) |
| 3 | $Ca_2+$ ion concentration in injecting fluid | (ppm) |
| 4 | $Mg_2+$ ion concentration in injecting fluid | (ppm) |
| 5 | $SO_4(2-)$ ion concentration in injecting fluid | (ppm) |
| 6 | $HCO_3-$ ion concentration in injecting fluid | (ppm) |
| 7 | TDS of injection fluid | (g/l) |
| 8 | pH | (-) |
| 9 | occurrence of scale formation | (TRUE/FALSE) |

**Table 3. The description of different parameters used for target functionality (III) - oil recovery**

| Item | Description (unit) | |
|---|---|---|
| 1 | Length of core | (mm) |
| 2 | Diameter of core | (mm) |
| 3 | Pore volume of core | (ml) |
| 4 | Porosity of the core | (-) |
| 5 | Sw, initial water saturation of core | (-) |
| 6 | So, initial oil saturation of core | (-) |
| 7 | Ko, relative permeability of oil in core | (mD) |
| 8 | Kw, relative permeability of water in core | (mD) |
| 9 | $CaCO_3$, weight percent of core | (-) |
| 10 | $SiO_2$, weight percent of core | (-) |
| 11 | $Al_2Si_2O_5(OH)_4$ weight percent of core | (-) |
| 12 | Acid number of oil | (mg KOH/g oil) |
| 13 | Specific gravity of oil | (-) |

| | | |
|---|---|---|
| 14 | API of oil | (-) |
| 15 | Asphaltene weight percent in oil | (-) |
| 16 | Viscosity of oil | (cp) |
| 17 | $C_2$ mole percent in oil composition | (-) |
| 18 | $C_3$ mole percent in oil composition | (-) |
| 19 | I-$C_4$ mole percent in oil composition | (-) |
| 20 | N-$C_4$ mole percent in oil composition | (-) |
| 21 | I-$C_5$ mole percent in oil composition | (-) |
| 22 | N-$C_5$ mole percent in oil composition | (-) |
| 23 | $C_6$ mole percent in oil composition | (-) |
| 24 | $C_7$ mole percent in oil composition | (-) |
| 25 | $C_8$ mole percent in oil composition | (-) |
| 26 | $C_9$ mole percent in oil composition | (-) |
| 27 | $C_{10}$ mole percent in oil composition | (-) |
| 28 | $C_{11}$ mole percent in oil composition | (-) |
| 29 | $C_{12}$+ mole percent in oil composition | (-) |
| 30 | $HCO_3$- ion concentration in formation water | (mol/l) |
| 31 | Cl- ion concentration in formation water | (mol/l) |
| 32 | $SO_4$(2-) ion concentration in formation water | (mol/l) |
| 33 | $Mg_2$+ ion concentration in formation water | (mol/l) |
| 34 | $Ca_2$+ ion concentration in formation water | (mol/l) |
| 35 | $SO_3$(2-) ion concentration in formation water | (mol/l) |
| 36 | $NO_2$- ion concentration in formation water | (mol/l) |
| 37 | $PO_4$(3-) ion concentration in formation water | (mol/l) |
| 38 | $Fe_2$+ ion concentration in formation water | (mol/l) |
| 39 | Na+ ion concentration in formation water | (mol/l) |
| 40 | K+ ion concentration in formation water | (mol/l) |
| 41 | Li+ ion concentration in formation water | (mol/l) |
| 42 | $Sr_2$+ ion concentration in formation water | (mol/l) |
| 43 | $Ba_2$+ ion concentration in formation water | (mol/l) |
| 44 | Ionic strength of formation water | (-) |
| 45 | TDS of formation water | (g/l) |
| 46 | $HCO_3$- ion concentration in injecting fluid | (mol/l) |
| 47 | Li+ ion concentration in injecting fluid | (mol/l) |
| 48 | K+ ion concentration in injecting fluid | (mol/l) |
| 49 | $Ca_2$+ ion concentration in injecting fluid | (mol/l) |
| 50 | $Mg_2$+ ion concentration in injecting fluid | (mol/l) |
| 51 | Na+ ion concentration in injecting fluid | (mol/l) |
| 52 | $SO_4$(2-) ion concentration in injecting fluid | (mol/l) |
| 53 | Cl- ion concentration in injecting fluid | (mol/l) |
| 54 | Ionic strength of injection fluid | (-) |
| 55 | TDS of injection fluid | (g/l) |
| 56 | Temperature | (C) |

## 3. Methods

Several methods have been employed in the present study, for regression as well as classification of parameters; including Bayesian Network (BN), Classification and Regression Trees (CART), Random Forest (RF) and Deep Neural Network (DNN). In order to keep the size of this manuscript within reasonable length, only an explanation of BN and RF methods is given in this section, which have outperformed the other applied techniques in terms of their established accuracy [31, 32].

### 3.1. Bayesian Network

A Bayesian network belongs to a class of graphical models, which concisely represent the probabilistic dependencies between a given set of (random) variables $X = \{X_1, X_2, \dots, X_N\}$, in the form of a directed acyclic graph (DAG). The DAG shapes such that its nodes represent the variables and its arrows represent probabilistic dependencies between the nodes. In this structure, an arrow goes from an influencing *parent* node to an influenced *child* node, in a one-directional way. Such a graphical structure enables estimation of the joint probability distribution. Provided that a variable only depends on its parent nodes, DAG defines a factorization of the joint probability distribution, for each variable, into a set of local probability distribution functions, in which the form of factorization is given by the Markov property of its network. The Markov chain framework considers the product of the conditional (local) probability distributions associated with each variable $X_i$, as the (global) joint probability distribution of variables in $X$ [33]. For the case of the factorization of the joint density function $f_X$ can be obtained by Nagarajan et al. (2013) [34]:

$$f_X(X) = \prod_{i=1}^{p} f_{X_i}(X_i \mid \Pi_{X_i}) \tag{1}$$

In which $\Pi_{X_i}$ represents the set of parents of $X_i$. For the random variables with discrete nature, the factorization of the joint probability distribution $P_X$ is obtained by:

$$P_X(X) = \prod_{i=1}^{p} P_{X_i}(X_i \mid \Pi_{X_i}) \tag{2}$$

For Each three disjoint subsets of nodes in DAG, say (A, B, C), a directed separation (*d-separation*) criterion is evaluated. Assuming node C to d-separate nodes A and B, then along every sequence of arcs between nodes A and B, there exists a node $v$, which either is positioned in C (not having any converging arcs) or has converging arcs (being pointed to along the network path by two arcs) and none of $v$ or the nodes that can be reached from $v$ (its descendants) are in C [34]. As the situation of a converging connection violates the d-separation criteria for the child node (Figure 2a), it can be assumed that the parent nodes (A and B) are not independent given the child node. As such, the Markov property stipulates:

$$P(A,B,C) = P(C \mid A,B)P(A)P(B) \tag{3}$$

For the other two scenarios – serial and diverging connections (Figures 2b and 2c) – the corresponding values measures as Equations 4 and 5, respectively:

$$P(A,B,C) = P(B \mid C)P(C \mid A)P(A) \tag{4}$$

$$P(A,B,C) = P(A \mid C)P(B \mid C)P(C) \tag{5}$$

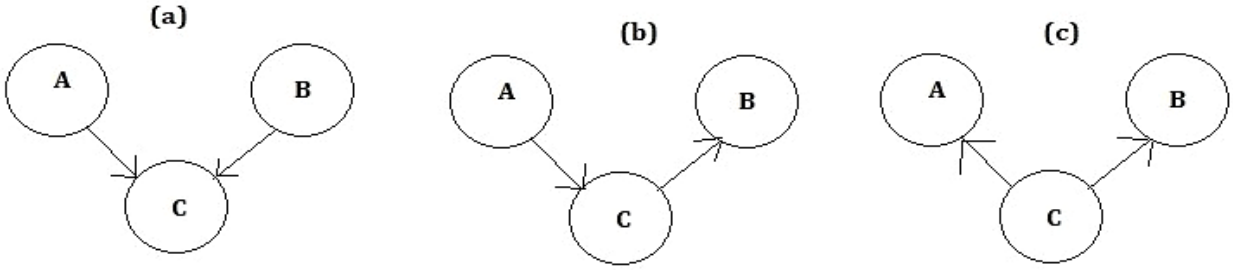



**Figure 2. The graphical separation for the converging connection (a), serial connection (b) and diverging connection (c) of fundamental connections in DAG**

The BN learning process aims to find an optimal structure in addition to its underlying parameters. In this respect, two approaches have been devised. The first approach analyzes the probabilistic relationships supervised by the Markov property of Bayesian networks with conditional independence tests and subsequently constructs a graph that satisfies the corresponding d-separation statements (Constraint-based algorithms). The other approach assigns a score to each BN candidate and maximizes it with a heuristic algorithm (Score-based algorithms) [34].

Once a Bayesian network has been established, approximate inference on an unknown value can be made by taking advantage of the BN's fundamental properties, which serves an added advantage of evading the curse of dimensionality, due to its mere usage of the local distributions [35]. In other words, the posterior probability of a target node can then be computed from the data generated by applying stochastic simulation to the distribution network, for a large number of cases. For this sake, two algorithms have been proposed – Logical Sampling (LS) and Likelihood weighting (LW). Traversing the nodes from the parent nodes down to children nodes, the LS algorithm generates a case by selecting values for each node –weighed by the probability of those values occurring at random. At each step, the weighing probability is either the prior or the conditional probability table entry for the sampled parent values. An instantiation of

all the nodes in the BN is later on created, once all the structure is visited. The collection of instantiation data enables estimation of the posterior probability for node $X$ given evidence. The LW algorithms works in a similar way to the former algorithm except that it adds the fractional likelihood of the evidence combination to the run count, instead of one [35].

### 3.2. Random Forest

The Random Forest is a class of ensemble learning techniques, which principally aggregates a collection of random decision trees. The individual trees are not necessarily optimal and are randomly perturbed. Such a diversity enables more extensive exploration of the tree predictors` space – enhancing the RF predictive performance. Each tree is composed of root, branch and leaf nodes, which is generated based on bootstrap sampling from the original training data. The optimal node splitting feature is selected, for each node of a tree, from a set of $n$ features, being randomly selected from a feature space of size $N$ [36]. If the number of features is less than the size of the feature space, the node splitting feature selection would decrease the correlation between different trees, which subsequently makes the average response of multiple regression trees to have lower expected variance than the individual regression trees. Nevertheless, an improvement in the predictive capability of the individual trees alongside increase in the number of selected features can result in an increase in the correlation between trees and therefore void any gains obtained from averaging multiple predictions.

Consider $X_{tr}(i,j)$ and $y(i)$ as being the training input feature $j$ and output response, respectively. For a sample $i$, $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, N$. The node splitting process would then attempt to select a feature $j$ from a set of $n$ features and partition the node $\chi_P$ into two child nodes with respect to a threshold $Z$. The child nodes – left and right – satisfy the conditions $X_{tr}(i \in \chi_P, j_s \leq Z)$ and $X_{tr}(i \in \chi_P, j_s > Z)$, respectively. Assume the node cost as being the sum of square deviances:

$$D(\chi_P) = \sum_{i \in \chi_P} (y(i) - \mu(\chi_P))^2 \tag{6}$$

where $\mu(\chi_P)$ denotes the expected value of the responses. Consequently, the objective function to optimize would be the reduction in cost for partition $\gamma$ at node $\chi_P$- the reward function (Equation 7).

$$\begin{aligned} C(\gamma, \chi_P) &= D(\chi_P) - D(\chi_L) - D(\chi_R) \\ \gamma^* &= \arg\max_{\gamma} C(\gamma, \chi_P) \end{aligned} \tag{7}$$

The optimal selection, $\gamma^* \in \chi_P$, maximizes the reward function. The node splitting process can be computationally expensive, as the complexity associated with each node split is of order $O(nm)$- requiring the checking of a total of $m$ partitions for a continuous feature with $m$ samples [36]. To deal with this complexity in the tree construction process, several recommendations have been proposed, such as applying the Principal Component Analysis (PCA) in the response matrix or using the basis functions to represent the response variables with the node cost [36]. The corresponding node cost functions to use in tree construction, in that case, would respectively take the form of Equations 8 to 9:

$$D(\chi_P) = \sum_{i \in \chi_P} (\varsigma(i) - \bar{\varsigma}(r))^T (\varsigma(i) - \bar{\varsigma}(r)) \tag{8}$$

$$D(\chi_P) = \sum_{i \in \chi_P} (c(i) - \mu_c(\chi_P))^T \Theta (c(i) - \mu_c(\chi_P)) \tag{9}$$

where $\varsigma(i)$ is the response obtained based on the principal components, $\bar{\varsigma}(r)$ is the principal components' mean vector, $c(i)$ is the vector of basis coefficients, $\mu_c(\chi_P)$ is the expected value of the basis coefficients vector and $\Theta$ represents the matrix of inner vector products.

The RF methodology relies on fitting the tree based on *bootstrap* samples from training data - $[(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)]$- while the randomized feature selection process is in effect. Assuming that $\bar{\gamma}(x, \Phi)$ represent the partition containing a test sample $x$ for the tree $\Phi$, the response of the tree can be obtained by Equation 10, with the corresponding weights, $w_i(x, \Phi)$, given by Equation 11 [36]:

$$y(x, \Phi) = \sum_{i=1}^{n} w_i(x, \Phi) y(i) \tag{10}$$

$$w_i(x, \Phi) = \frac{1_{\{x_{tr}(i) \in \bar{\gamma}(x, \Phi)\}}}{\#\{r : x_{tr}(i) \in \bar{\gamma}(x_{tr}(r), \Phi)\}} \tag{11}$$

Should a collection of a number of $T$ trees be accumulated - $\Phi_1, \Phi_2, \dots, \Phi_T$ - the average RF prediction for the test sample can be obtained by incorporating the average weights over the forest:

$$w_i(x) = \frac{1}{T} \sum_{j=1}^{T} w_i(x, \Phi_j) \tag{12}$$

$$\hat{y}(x) = \sum_{i=1}^{n} w_i(x) y(i) \tag{13}$$

## 4. Results

The results obtained are applicable to three set of target functionalities, based on which the data was originally established. In addition, the results pinpoint to a performance benchmark amongst different methods applied in selected target functionalities. Table 4 lists the results obtained for the importance rank of influencing parameters related to target functionality (I), obtained by the RF method. The results are valid for analyzing permeability impairment in sandstone matrix. In this setting, the ions in the formation water have shown an influencing rank in the order of $SO_4(2-)$>$Ba_2+$> $Sr_2+$>$Ca_2+$. In essence, two sets of parameters are enlisted in this table; namely, ion-related parameters (micro-scale) as well as the parameters related to the physics of the field (such as initial permeability) or experimental conditions devised (macro-scale). As evident from the table, both micro-scale and macro-scale parameters have driven an influencing role in the permeability impairment process; nevertheless, the importance of macro-scale parameters have ranked higher. Figure 3 depicts the underlying Bayesian network of influencing parameters related to target functionality (I) for the same set of data. The figure is informative, as it provides the first illustration of the exact interplay amongst different parameters in the data on target functionality (I). Only the statistically-significant arcs have been drawn. It is noticeable that the earlier conclusion made on the microscale/macroscale parameters importance comparison is also confirmed by the Bayesian network – placing mostly macroscale parameters in the parent nodes. The results presented in Table 1 is further confirmed by the applying the tree pruning to the data (Figure 4). As can be seen, the highly-influencing parameters in target functionality (I) – such as pore volume, initial permeability or $SO_4(2-)$ concentration- are detected with higher splitting position in the optimally-pruned tree for data in target functionality (I). The numbers at the target end of the branches (i.e. end lines) in Figure 4, refer to the number of cases received in that end terminal.

**Table 4. The importance rank of influencing parameters related to target functionality (I) in sandstone matrix**

| Importance rank | Parameter |
|---|---|
| 1 | pore volume |
| 2 | Initial permeability |
| 3 | $SO_4(2-)$ ion concentration in formation water |
| 4 | Rate of injection |
| 5 | $Ba_2+$ ion concentration in formation water |
| 6 | Pressure difference along the core |
| 7 | $Sr_2+$ ion concentration in formation water |
| 8 | Temperature |
| 9 | $Ca_2+$ ion concentration in formation water |

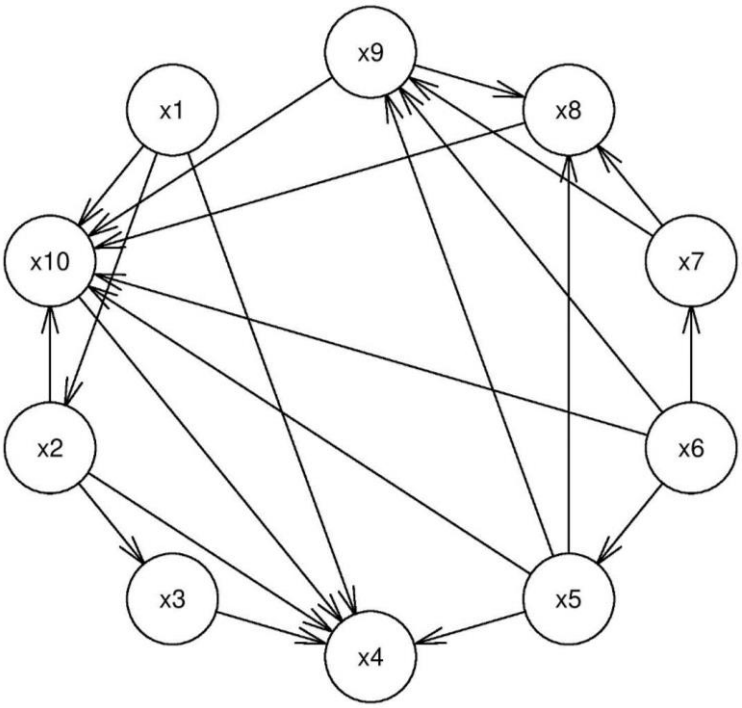



**Figure 3. The underlying Bayesian network of influencing parameters related to target functionality (I) in sandstone matrix**

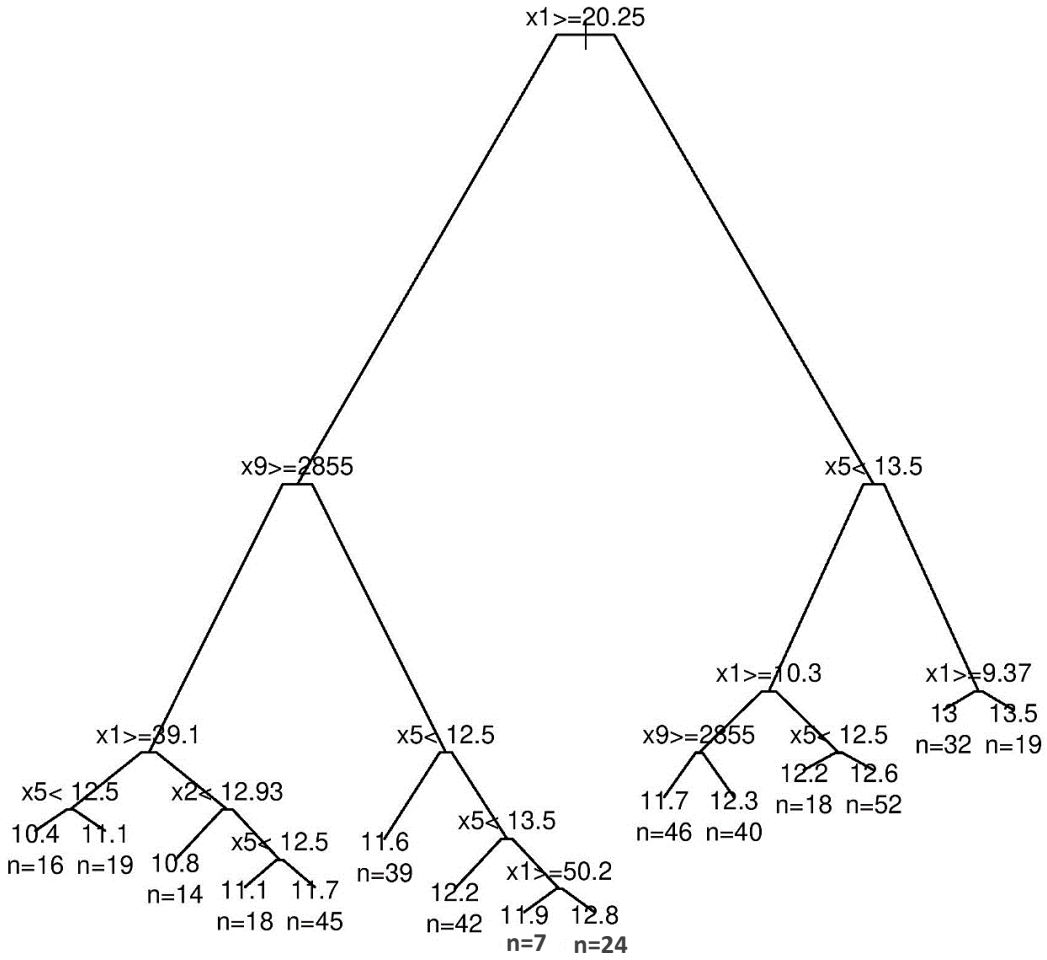



**Figure 4. The optimally-pruned tree for data in target functionality (I)**

Using the improved knowledge of parameters` roles mentioned above, this work attempted to improve on the predictive ability in target functionality (I). Figure 5 shows our results for the final permeability in sandstone matrix, obtained by the RF method. The results pertain to bootstrapped ones with a fraction of 70% for training, which attest to the accuracy of the method. As evident, our results have outperformed the competing results of the gene expression method of Rostami et al. [25]. Our methodology is also competitive to other hybrid scheme prepositions for the same data [23, 24] – yielding a sum of error squared ($R^2$) of 0.987 and 0.978 on bootstrapped results for the RF and BN methods, respectively. It should be noticed that the mentioned hybrid machine learning attempts on the same data, merely report their $R^2$ measurements on the whole data set, and not on a bootstrapped sample, which would have been different if tested otherwise.

Table 5 provides a performance benchmark of different machine learning algorithms analyzed for target functionality (I) – reporting the mean error percentile of bootstrapped results in each case. For the deep machine learning case, the $H_2O$ AutoML scheme was used [37], which allows for automatic inspection of over 270 neural network models for optimal detection. Based on the results, the RF method provides the most accurate output for target functionality (I), with a mean error percentile of less than 5% on the bootstrapped results.

**Table 5. The mean error percentile of bootstrapped results of different machine learning algorithms for target functionality (I)**

| Method | Mean error (%) |
|---|---|
| Recursive partitioning | 0.10 |
| Random Forest | 0.05 |
| Bayesian Network | 0.11 |
| Deep Neural Network | 0.14 |

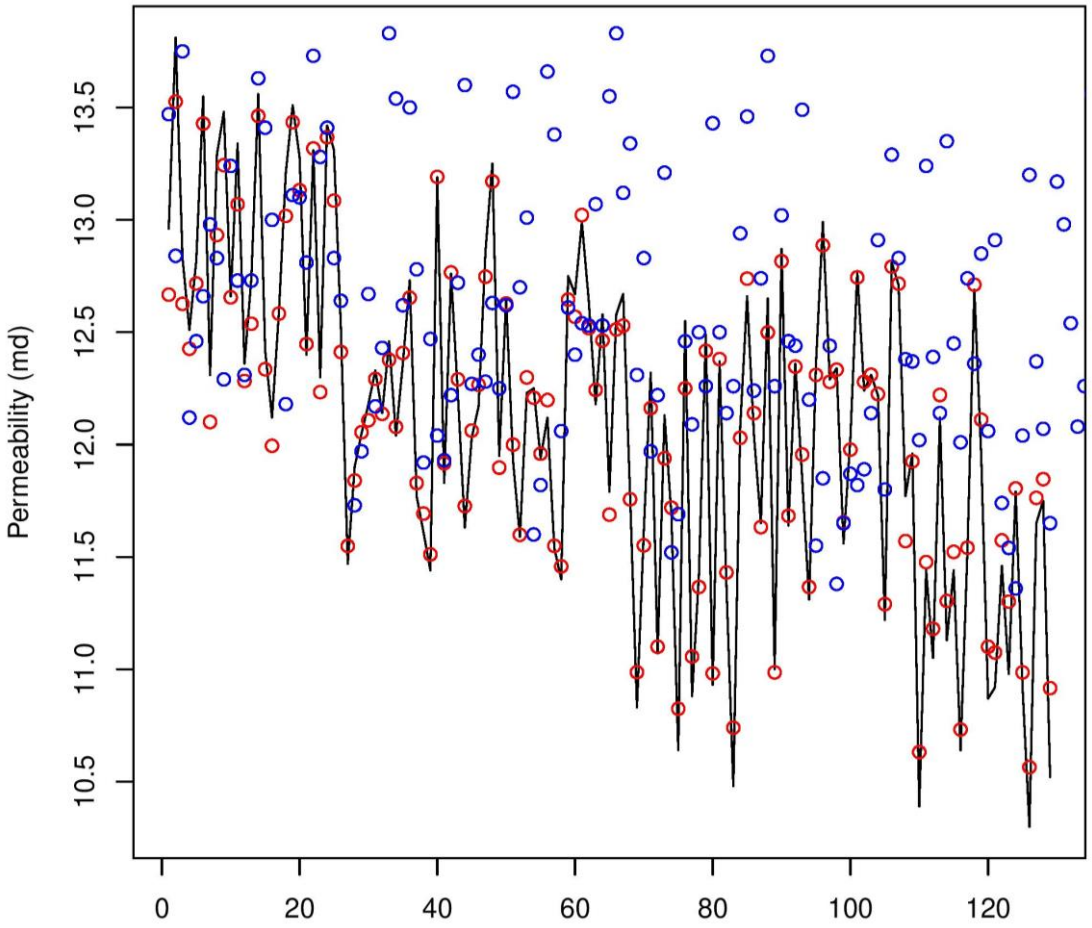



**Figure 5. The bootstrapped final permeability results for target functionality (I). The experimental data (solid line), the Gene Expression Programming results of Rostami et al. [25] (blue circles) and the random forest results of the present work (red circles).**

In an analogous way, the data related to target functionality (II) were analyzed, based on which the importance rank of influencing parameters were ascertained (Table 6). The nature of the data related to this functionality had been of the categorical type (i.e. TRUE/FALSE), which referred to the occurrence of calcium carbonate scale formation in the field [4]. The data was also more limited than the data in the other two functionalities, in terms of the number of macro-scale parameters enlisted - mostly considering ions data for a practical purpose. However, the results again indicate the highest-ranking parameter in the group as being of a macro-scale type (i.e. the "field" parameter). As determined by the RF method, the hierarchy of the ion effects in the process has been in the order of $Ca_2$+>Na+>$Mg_2$+>$HCO_3$->$SO_4$(2-) in the injecting fluid, while placing the pH parameter in the least influencing rank. Table 7 lists the performance of different machine learning algorithms for target functionality (II). Given the categorical type of output, there performance is reported based on the accuracy of the confusion matrix for bootstrapped results. Both RF and BN methods have outperformed the other machine learning methods, in terms of their accuracy. This performance comparison would make the RF results more trustable, including the deductions made earlier on the ranking hierarchy of the ions.

**Table 6. The importance rank of influencing parameters related to target functionality (II)**

| Importance rank | Parameter |
|---|---|
| 1 | Field |
| 2 | $Ca_2$+ ion concentration in injecting fluid |
| 3 | TDS of injection fluid |
| 4 | Na+ ion concentration in injecting fluid |
| 5 | $Mg_2$+ ion concentration in injecting fluid |
| 6 | $HCO_3$,- ion concentration in injecting fluid |
| 7 | $SO_4$(2-) ion concentration in injecting fluid |
| 8 | pH |

**Table 7. The confusion matrix's accuracy of bootstrapped results obtained from different machine learning algorithms for target functionality (II)**

| Method | Accuracy (%) |
|---|---|
| Recursive partitioning | 80 |
| Random Forest | 88 |
| Bayesian Network | 91 |
| Deep Neural Network | 70 |

Table 8 lists the RF results obtained for the importance rank of influencing parameters related to target functionality (III). As evident, the highest-ranking parameter in the list is detected with a macro-scale type, in the carbonate matrix. Since the data accounted for the source of ion introduction – injecting or formation – it was possible to determine on the importance rank of each ion species, based on its introduction source, which is the first analysis of its kind, to the knowledge of the authors. The dataset used for target functionality (III) in the present article cannot be distributed due to the licensing issues.

Based on the RF results (Table 8), the effective ions have shown a different influencing hierarchy compared to the other sectors studied. For a given type of ion, the influencing rank on oil recovery has also been different based on the introduction source into the carbonate matrix. For instance, the importance level of HCO3- ion in the formation water has been higher than that in the injecting water. For some other ions, such as $SO_4$(2-), the injecting water content has rendered more important than the content in the formation water. This finding potentially reveals a more complex phenomenon underlying the oil recovery process, which requires further investigation. On the other hand, the overall effectiveness of ions shows an altered arrangement compared to the other carbonate case (item II), which nullifies any general deductions on the global effectiveness of an ion over the others – suggesting its case dependency.

For instance, in a low salinity water injection context, a recent research [38] reports that neither the cation/$SO_4$(2-) concentration nor the difference of sulfate ion concentration change in brines show effectiveness towards tertiary oil recovery. The data mining results for target functionality (III) indicate that the most influential parameter in the list is of a macro-scale type, which sustains through all the functionalities/environments studied.

**Table 8. The importance rank of influencing parameters related to target functionality (III) in carbonate matrix**

| Importance rank | Parameter |
|---|---|
| 1 | Temperature |
| 2 | Specific gravity of oil |
| 3 | Porosity |
| 4 | $HCO_3$- ion concentration in formation water |
| 5 | Na+ ion concentration in formation water |
| 6 | $SO_4$(2-) ion concentration in injecting fluid |
| 7 | TDS of injection fluid |
| 8 | $C_l$- ion concentration in injecting fluid |
| 9 | $Ca_2$+ ion concentration in injecting fluid |
| 10 | Pore volume of core |
| 11 | Ionic strength of injection fluid |
| 12 | Asphaltene weight percent in oil |
| 13 | Na+ ion concentration in injecting fluid |
| 14 | $CaCO_3$ weight percent of core |
| 15 | $HCO_3$- ion concentration in injecting fluid |
| 16 | API of oil |
| 17 | Sw, initial water saturation of core |

| | |
|---|---|
| 18 | Diameter of core |
| 19 | K+ ion concentration in injecting fluid |
| 20 | So, initial oil saturation of core |
| 21 | Ko, relative permeability of oil in core |
| 22 | $SiO_2$ weight percent of core |
| 23 | $Mg_2$+ ion concentration in injecting fluid |
| 24 | Length of core |
| 25 | Viscosity of oil |
| 26 | Acid number of oil |
| 27 | TDS of formation water |
| 28 | K+ ion concentration in formation water |
| 29 | Kw, relative permeability of water in core |
| 30 | $SO_4$(2-) ion concentration in formation water |
| 31 | Ionic strength of formation water |
| 32 | $Al_2Si_2O_5(OH)_4$ weight percent of core |
| 33 | Li+ ion concentration in injecting fluid |
| 34 | $Ca_2$+ ion concentration in formation water |
| 35 | $Mg_2$+ ion concentration in formation water |
| 36 | $PO_4$(3-) ion concentration in formation water |
| 37 | Li+ ion concentration in formation water |
| 38 | $NO_2$- ion concentration in formation water |
| 39 | $Sr_2$+ ion concentration in formation water |
| 40 | $Fe_2$+ ion concentration in formation water |
| 41 | $C_8$ mole percent in oil composition |
| 42 | $C_3$ mole percent in oil composition |
| 43 | I-$C_4$ mole percent in oil composition |
| 44 | $C_7$ mole percent in oil composition |
| 45 | $C_l$- ion concentration in formation water |
| 46 | I-$C_5$ mole percent in oil composition |
| 47 | $SO_3$(2-) ion concentration in formation water |
| 48 | $C_2$ mole percent in oil composition |
| 49 | $Ba_2$+ ion concentration in formation water |
| 50 | $C_{12}$+ mole percent in oil composition |
| 51 | N-$C_5$ mole percent in oil composition |
| 52 | $C_6$ mole percent in oil composition |
| 53 | $C_{10}$ mole percent in oil composition |
| 54 | N-$C_4$ mole percent in oil composition |
| 55 | $C_{11}$ mole percent in oil composition |
| 56 | $C_9$ mole percent in oil composition |

## 5. Conclusion

The data mining results indicate a rolling importance for the confluent effect of considered ion species, which alters under different environments. This essentially rejects the prior propositions on the existence of a global order list for the effectiveness of ions for selected functionalities. For the carbonate matrix, the random forest results clearly distinguish between the source of ion introduction into the matrix—injecting or formation water—and its importance rank in the sequel, for oil recovery purposes. This latter conclusion is notable as it provides the first quantitative confirmation of the source of ion significance towards its overall functionality, bringing a novel concept to the field. For all the target functionalities/matrix environments studied, the most influential parameter was detected as being of a macro-scale type, which does not include an ion. In other words, in neither of the cases studied, an ion was ranked as the most influential parameter in the list. The minimal errors obtained over the bootstrapped results indicate that the machine learning methodologies applied have been successful in capturing the experimental/field data within major rock types (carbonate and sandstone) over studied functionalities. The random forest and Bayesian network methods stand out as the most accurate techniques amongst the other machine learning strategies applied to the sandstone case, whose results outperform the most recent hybrid method predictions on the same data.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization, B.F., M.M. and P.P.; methodology, B.F., M.M. and B.F.; validation, B.F., M.M. and P.P.; formal analysis, B.F., M.M., and P.P.; investigation, B.F. and M.M.; writing—original draft preparation, B.F.; writing—review and editing, B.F., M.M. and P.P.; visualization, B.F. All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The data presented in this study are available in article.

### 6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 6.4. 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.

## 7. References

[1] Frenier, W. W., & Ziauddin, M. (2008). Formation, removal, and inhibition of inorganic scale in the oilfield environment (p. 808). Richardson, TX: Society of Petroleum Engineers, Texas, United States.

[2] Hajirezaie, S., Wu, X., & Peters, C. A. (2017). Scale formation in porous media and its impact on reservoir performance during water flooding. Journal of Natural Gas Science and Engineering, 39, 188–202. doi:10.1016/j.jngse.2017.01.019.

[3] Azizi, J., Shadizadeh, S. R., Khaksar Manshad, A., & Jadidi, N. (2018). Effects of pH and temperature on oilfield scale formation. Iranian Journal of Oil and Gas Science and Technology, 7(3), 18-31. doi:10.22050/ijogst.2017.58038.1350.

[4] Al-Hajri, N. M., Al-Ghamdi, A., Tariq, Z., & Mahmoud, M. (2020). Scale-Prediction/Inhibition Design Using Machine-Learning Techniques and Probabilistic Approach. SPE Production & Operations, 35(04), 0987–1009. doi:10.2118/198646-pa.

[5] Kelland, M. A. (2011). Effect of Various Cations on the Formation of Calcium Carbonate and Barium Sulfate Scale with and without Scale Inhibitors. Industrial & Engineering Chemistry Research, 50(9), 5852–5861. doi:10.1021/ie2003494.

[6] Zhang, P., Zhang, Z., Liu, Y., Kan, A. T., & Tomson, M. B. (2019). Investigation of the impact of ferrous species on the performance of common oilfield scale inhibitors for mineral scale control. Journal of Petroleum Science and Engineering, 172, 288–296. doi:10.1016/j.petrol.2018.09.069.

[7] Ahmadi, M. A. (2015). Developing a Robust Surrogate Model of Chemical Flooding Based on the Artificial Neural Network for Enhanced Oil Recovery Implications. Mathematical Problems in Engineering, 1–9. doi:10.1155/2015/706897.

[8] Baghernezhad, D., Siavashi, M., & Nakhaee, A. (2019). Optimal scenario design of steam-assisted gravity drainage to enhance oil recovery with temperature and rate control. Energy, 166, 610–623. doi:10.1016/j.energy.2018.10.104.

[9] Giro, R., Filho, S., Ferreira, R., Engel, M., & Steiner, M.B. (2019) Offshore Technology Conference Brasil - Artificial Intelligence-Based Screening of Enhanced Oil Recovery Materials for Reservoir-Specific Applications. Offshore Technology Conference Offshore Technology Conference Brasil - Rio de Janeiro, Brazil, October 31. doi:10.4043/29754-ms.

[10] Siavashi, M., & Doranehgard, M. H. (2017). Particle swarm optimization of thermal enhanced oil recovery from oilfields with temperature control. Applied Thermal Engineering, 123, 658–669. doi:10.1016/j.applthermaleng.2017.05.109.

[11] Vo Thanh, H., Sugai, Y., & Sasaki, K. (2020). Application of artificial neural network for predicting the performance of CO2 enhanced oil recovery and storage in residual oil zones. Scientific Reports, 10(1). doi:10.1038/s41598-020-73931-2.

[12] Cheraghi, Y., Kord, S., & Mashayekhizadeh, V. (2021). Application of machine learning techniques for selecting the most suitable enhanced oil recovery method; challenges and opportunities. Journal of Petroleum Science and Engineering, 205, 108761. doi:10.1016/j.petrol.2021.108761.

[13] You, J., Ampomah, W., & Sun, Q. (2020). Development and application of a machine learning based multi-objective optimization workflow for CO2-EOR projects. Fuel, 264, 116758. doi:10.1016/j.fuel.2019.116758.

[14] Koroteev, D., & Tekic, Z. (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. Energy and AI, 3, 100041. doi:10.1016/j.egyai.2020.100041.

[15] Dias, L. O., Bom, C. R., Faria, E. L., Valentín, M. B., Correia, M. D., de Albuquerque, M. P., … Coelho, J. M. (2020). Automatic detection of fractures and breakouts patterns in acoustic borehole image logs using fast-region convolutional neural networks. Journal of Petroleum Science and Engineering, 191, 107099. doi:10.1016/j.petrol.2020.107099.

[16] Nozohour-leilabady, B., & Fazelabdolabadi, B. (2016). On the application of artificial bee colony (ABC) algorithm for optimization of well placements in fractured reservoirs; efficiency comparison with the particle swarm optimization (PSO) methodology. Petroleum, 2(1), 79–89. doi:10.1016/j.petlm.2015.11.004.

[17] Wood, D. A. (2016). Metaheuristic profiling to assess performance of hybrid evolutionary optimization algorithms applied to complex wellbore trajectories. Journal of Natural Gas Science and Engineering, 33, 751–768. doi:10.1016/j.jngse.2016.05.041.

[18] Hassan, A., Elkatatny, S., & Abdulraheem, A. (2019). Application of Artificial Intelligence Techniques to Predict the Well Productivity of Fishbone Wells. Sustainability, 11(21), 6083. doi:10.3390/su11216083.

[19] Panja, P., Velasco, R., Pathak, M., & Deo, M. (2018). Application of artificial intelligence to forecast hydrocarbon production from shales. Petroleum, 4(1), 75–89. doi:10.1016/j.petlm.2017.11.003.

[20] Alatrach, Y., Mata, C., Omrani, P.J., Saputelli L., Narayanan R., & Hamdan M. (2020) Prediction of Well Production Event Using Machine Learning Algorithms. Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE.

[21] Bikmukhametov, T., & Jäschke, J. (2019). Oil Production Monitoring using Gradient Boosting Machine Learning Algorithm. IFAC-PapersOnLine, 52(1), 514–519. doi:10.1016/j.ifacol.2019.06.114.

[22] Hosseini-Dastgerdi, Z., & Jafarzadeh-Ghoushchi, S. (2019) Investigation of Asphaltene Precipitationusing Response Surface Methodology Combined with Artificial Neural Network. Journal of Chemical and Petroleum Engineering 53(2) 153-167. doi:10.22059/jchpe.2019.261438.1238.

[23] Ahmadi, M. A., Mohammadzadeh, O., & Zendehboudi, S. (2017). A cutting edge solution to monitor formation damage due to scale deposition: Application to oil recovery. The Canadian Journal of Chemical Engineering, 95(5), 991–1003. doi:10.1002/cjce.22776.

[24] Ahmadi, M., & Chen, Z. (2020). Machine learning-based models for predicting permeability impairment due to scale deposition. Journal of Petroleum Exploration and Production Technology, 10(7), 2873–2884. doi:10.1007/s13202-020-00941-1.

[25] Rostami, A., Shokrollahi, A., Shahbazi, K., & Ghazanfari, M. H. (2019). Application of a new approach for modeling the oil field formation damage due to mineral scaling. Oil & Gas Science and Technology – Revue d'IFP Energies Nouvelles, 74, 62. doi:10.2516/ogst/2019032.

[26] Li, H., Yu, H., Cao, N., Tian, H., & Cheng, S. (2020). Applications of Artificial Intelligence in Oil and Gas Development. Archives of Computational Methods in Engineering, 28(3), 937–949. doi:10.1007/s11831-020-09402-8.

[27] Ahmadi, S., Hosseini, M., Tangestani, E., Mousavi, S. E., & Niazi, M. (2020). Wettability alteration and oil recovery by spontaneous imbibition of smart water and surfactants into carbonates. Petroleum Science, 17(3), 712–721. doi:10.1007/s12182-019-00412-1.

[28] Fathi, S. J., Austad, T., & Strand, S. (2010). "Smart Water" as a Wettability Modifier in Chalk: The Effect of Salinity and Ionic Composition. Energy & Fuels, 24(4), 2514–2519. doi:10.1021/ef901304m.

[29] Montazeri, M., Fazelabdolabadi, B., Shahrabadi, A., Nouralishahi, A., HallajiSani, A., & Moosavian, S. M. A. (2020). An experimental investigation of smart-water wettability alteration in carbonate rocks – oil recovery and temperature effects. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 1–13. doi:10.1080/15567036.2020.1759735.

[30] Zhao, L., & Zhang, B. (2020) Measurement and correlation of solubility of 2-chloro-3-(trifluoromethyl)pyridine in pure solvents and ethanol + n-propanol mixtures. Journal of Molecular Liquids 298(15), 112103. doi:10.1016/j.molliq.2019.112103.

[31] Genuer, R., & Poggi, J. M. (2020). Random Forests with R. Springer, Germany. doi:10.1007/978-3-030-56485-8.

[32] Kelleher, J.D. (2019). Deep Learning. The MIT Press, Massachusetts, United States.

[33] Korb, K. B., & Nicholson, A. E. (2010). Bayesian artificial intelligence. CRC Press, Florida, United States.

[34] Nagarajan, R., Scutari, M., & Lebre, S. (2013). Bayesian Networks in R with Applications in System Biology. Springer, New York, United States. doi:10.1007/978-1-4614-6446-4.

[35] Fazelabdolabadi, B., & Golestan, M. H. (2020). Towards Bayesian Quantification of Permeability in Micro-scale Porous Structures – The Database of Micro Networks. HighTech and Innovation Journal, 1(4), 148–160. doi:10.28991/hij-2020-01-04-02.

[36] Rahman, R., Dhruba, S. R., Ghosh, S., & Pal, R. (2019). Functional random forest with applications in dose-response predictions. Scientific Reports, 9(1). doi:10.1038/s41598-018-38231-w.

[37] LeDell, E., & Poirier, S. (2020) H2O AutoML: Scalable Automatic Machine Learning. 7th ICML Workshop on Automated Machine Learning (AutoML), Virtual workshop.

[38] Salimova, R. (2021). Data-driven analysis of low salinity waterflooding in carbonates. M.Sc. Dissertation, Nazarbayev University, Astana, Kazakhstan.