scientific reports

Check for updates

OPEN

# Application of XGBoost model for in-situ water saturation determination in Canadian oil-sands by LF-NMR and density data

Strahinja Markovic[1,2]✉, Jonathan L. Bryan[4], Reza Rezaee[2], Aman Turakhanov[1], Alexey Cheremisin[1], Apostolos Kantzas[3] & Dmitry Koroteev[1]

**Water saturation determination is among the most challenging tasks in petrophysical well-logging, which directly impacts the decision-making process in hydrocarbon exploration and production. Low-field nuclear magnetic resonance (LF-NMR) measurements can provide reliable evaluation. However, quantification of oil and water volumes is problematic when their NMR signals are not distinct. To overcome this, we developed two machine learning frameworks for predicting relative water content in oil-sand samples using LF-NMR spin–spin ($T_2$) relaxation and bulk density data to derive a model based on Extreme Gradient Boosting. The first one facilitates feature engineering based on empirical knowledge from the $T_2$ relaxation distribution analysis domain and mutual information feature extraction technique, while the second model considers whole samples' NMR $T_2$-relaxation distribution. The NMR $T_2$ distributions were obtained for 82 Canadian oil-sands samples at ambient and reservoir temperatures (164 data points). The true water content was determined by Dean-Stark extraction. The statistical scores confirm the strong generalization ability of the feature engineering LF-NMR model in predicting relative water content by Dean-Stark—root-mean-square error of 0.67% and mean-absolute error of 0.53% ($R^2 = 0.90$). Results indicate that this approach can be extended for the improved in-situ water saturation evaluation by LF-NMR and bulk density measurements.**

**Abbreviations**

| | |
|---|---|
| LF-NMR | Low-field nuclear magnetic resonance |
| OOIP | Original oil in place |
| EOR | Enhanced oil recovery |
| TE | Echo spacing |
| RMSE | Root mean square error |
| MAE | Mean absolute error |
| XGB | Extreme gradient boosting |
| FE | Feature engineering |
| FS | Full spectrum |
| XGB-FE | Extreme gradient boosting model based on feature engineering |
| XGB-FS | Extreme gradient boosting model based on full $T_2$ distribution (spectrum) |
| DS | Dean-Stark |
| DS-w | Water content by Dean-Stark |
| L1 | Lasso regression |
| L2 | Ridge regression |
| CPMG | Carr-Purcell-Meiboom-Gill pulse sequence |
| SNR | Signal-to-noise ratio |

[1]Centre for Petroleum Science and Engineering, Skolkovo Institute of Science and Technology, Sikorsky Street 11, Moscow, Russian Federation 121205. [2]Curtin University, Kent Street, Perth, Bentley, WA 6845, Australia. [3]University of Calgary, Calgary, AB, Canada. [4]PERM Inc., Calgary, AB, Canada. ✉email: strahinja.markovic@postgrad.curtin.edu.au

| | |
|---|---|
| CANOVA | Continuous analysis of covariance |
| MI | Mutual information |
| BO | Bayesian optimization |
| LOOCV | Leave-one-out cross-validation |
| X-ray CT | X-ray computed tomography |
| BSS | Blind-source signal separation |
| ANNs | Artificial neural networks |

The ever-growing energy demand and volatile oil prices are driving the development of new technologies to optimize hydrocarbon production and increase recoverable reserves from conventional and unconventional deposits. To better understand their amount, petrophysical surveying is regularly employed for quantifying the in-situ oil and water saturations in reservoirs. Determining water saturation is one of the most challenging tasks involving many techniques that typically produce significantly different estimates of original oil in place (OOIP)[1]. In unconventional reservoirs, the accuracy of these estimations becomes even more important as it directly impacts the development of enhanced oil recovery (EOR) strategy. The same applies to oil-sands bitumen deposits, which are estimated to account for 30% of total world reserves[2].

Conventionally, in-situ saturations of fluids are determined through resistivity logging. However, due to the mixed or low salinity of formation water that may be present in shallower parts of the oil-sands reservoir, the water saturation volumes can be over or underestimated[1]. The low-field nuclear magnetic resonance (LF-NMR) logging tools have proved to be a valuable alternative since the measurements are non-invasive and independent of lithology. In the past 20 years, NMR measurements were used for water–oil emulsions characterization[3] and in recent years for the in-situ fluid saturation determination[4,5]. Although the application of NMR measurements have been demonstrated to be successful for emulsion characterization, they are limited to emulsions with high water content (> 10 wt%), and it usually requires reference measurements of oil and water NMR amplitudes necessary for obtaining additional parameters such as relative hydrogen index[3,6]. Furthermore, the application of NMR emulsion models in well-logging is even more problematic due to the alteration of NMR relaxation behavior of fluids confined in the reservoir pore space. One of the well-known approaches that had considerable success in addressing this effect is based on NMR spin–spin relaxation ($T_2$) distribution peak deconvolution[7]. The concept behind this approach is that viscous bitumen relaxes faster in an NMR distribution even than surface bound water, so early $T_2$ signals are attributed to in-situ bitumen, and later $T_2$ signals correspond to the water saturation in the rock. Unfortunately, in cases when the NMR $T_2$ distribution contain signals from bitumen, water films on grains surrounded by bitumen or heavy oils, and fast-relaxing clay bound water, a significant overlap of the signals occurs, making their separation challenging. Additionally, this approach requires not only the determination of water and oil NMR amplitudes but also the independent measurement of their volume or mass. Alternative methods involve 2D LF-NMR measurements, where instead of using one NMR parameter (i.e. $T_2$ relaxation), additional parameters are employed (i.e. $T_1$ relaxation or diffusion) to obtain so-called 2D NMR maps[8–10] which can theoretically help to separate these overlapping bitumen and water signals. Application of 2D maps showed considerable success in fluid saturation evaluation, compared to 1D $T_2$ relaxation distribution analysis, since $T_1$ relaxation or diffusion of reservoir fluids can be sufficiently different, thus enabling relatively simple separation of their signals. However, 2D NMR is slower and more expensive to run, and there can still be instances where these signals are not distinct, in which case estimation of fluid types and fluid volumes can be challenging and require advanced analysis involving blind-source signal separation (BSS), clustering algorithms, and a certain degree of knowledge in 2D NMR maps interpretation[5].

In the past few years, a noticeable surge of machine learning applications in the petrophysical well-logging has been seen, ranging from improved interpretation of logs[11] and forecasting of stress in horizontal wells[12] to the utilization of neural networks to synthesize the artificial NMR $T_2$ logs[13]. In this study, two machine learning techniques are employed to improve the water content determination in oil-sands, using as inputs the $T_2$ relaxation and bulk density data, along with Extreme Gradient Boosting (XGBoost) algorithm. The first modeling approach is based on a feature engineering process that reduces the number of inputs while maximizing model generalization capacity. This was achieved by deriving new features using empirical knowledge from the $T_2$ distribution analysis domain and a feature extraction technique based on information theory. In contrast, the second approach considers as an input the whole NMR $T_2$ distribution of the sample, aiming to preserve all available information originating from fluids residing in the sample pore space. The dataset comprised 82 oil-sands core samples recovered from northern Alberta in Canada. Water content percentage relative to of the total mass of the sample was determined by Dean-Stark extraction (%DS-w). The model training and prediction test scores of the models were evaluated using three statistical metrics and a leave-one-out cross-validation (LOOCV). These scores were compared with water content predictions based on the previously published deconvolution approach[7].

## Theory

**LF-NMR measurements for water saturation determination.** LF-NMR logging tools are measuring the response of hydrogen protons in fluids rich in hydrogen, such as oils, bitumen and water. As these tools are primarily sensitive to liquids, the response from solids (reservoir rock) remains invisible in the LF-NMR $T_2$-relaxation distribution. As the physicochemical properties of these liquids vary, their relaxation will vary accordingly, thus allowing inference of their properties. LF-NMR measurements are performed in two steps: first, the NMR probe introduces the external magnetic field which polarizes the H protons ($T_1$-relaxation); and second, the subsequent series of short radio-frequency pulses are applied in order to produce the signal decay curve, which represents the relaxation of H protons returning to the previous state ($T_2$-relaxation). After mathematical inversion of decay curves, the $T_2$ relaxation can be represented time-domain. Three main processes

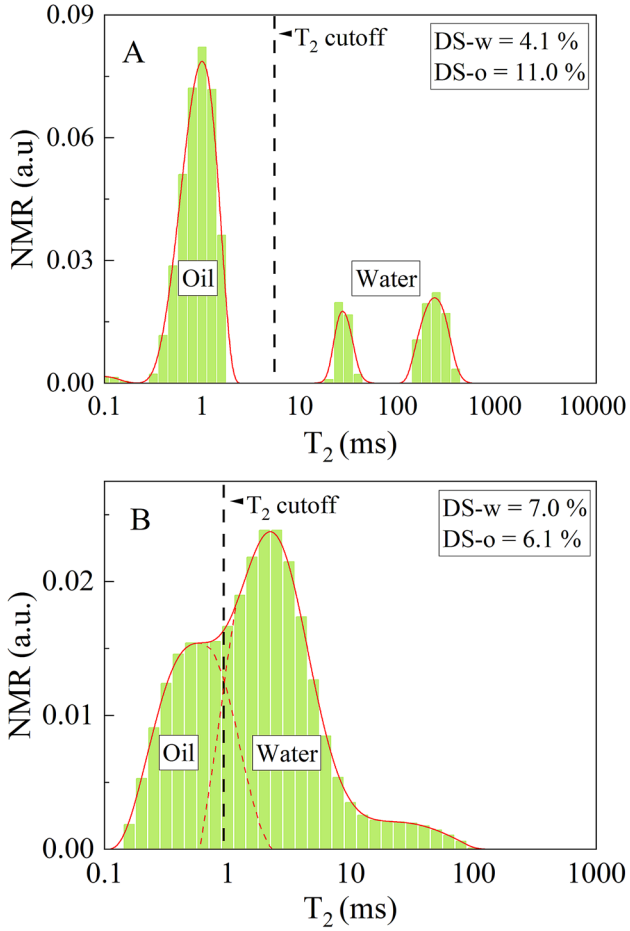


**Figure 1.** Representative NMR $T_2$ distributions of two oil-sand samples. (**A**) An example of distinct oil and water signals where a simple cutoff method can be used for oil–water separation. (**B**) An example of NMR $T_2$ distribution with overlapped oil and water signals where deconvolution with $T_2$ cutoff cannot provide a satisfactory solution. Black vertical dashed lines present potential cutoff times. DS-w and DS-o are percentages of water and oil by Dean-Stark, respectively, relative to solids.

comprise the total $T_2$-relaxation including bulk relaxation, surface relaxation, and diffusion relaxation due to the gradient in a magnetic field. In this work, the benchtop LF-NMR relaxometer was used in which the gradient is absent, thus the diffusion term can be neglected.

$$\frac{1}{T_2} = \frac{1}{T_{2B}} + \frac{1}{T_{2S}} \tag{1}$$

$$\frac{1}{T_{2s}} = \rho_2 \left(\frac{S}{V}\right) \tag{2}$$

$T_{2B}$ represents the relaxation occurring in bulk fluids or fluids in large pores, $T_{2S}$ quantifies the relaxation of fluids in smaller pores. Also, $\rho_2$ is $T_2$ surface relaxivity, $S/V$ is a ratio of the fluid volume and surface of the pore. Each of these mechanisms will contribute to the total relaxation in varied proportions depending on reservoir rock properties and physicochemical properties of the fluids such as rock wettability, pore size and pore surface area, fluid viscosity and chemical composition.

Assuming that the oil-sands are largely water-wet, water will generally be found in the corners of connected sand grains and potentially also as a thin film over the grain surface. The principal relaxation mechanism of hydrogen protons in high viscosity oils and bitumen would be bulk relaxation, while water would be under the strong influence of surface relaxation, and with bulk relaxation playing a smaller role in the water $T_2$ values. Bulk relaxation and surface relaxation times of water and oils are unique for the most part, that is, the oil molecules relax generally quicker relative to the water molecules. When the NMR $T_2$ distribution contains discrete responses of oil and water (Fig. 1A), a simple cutoff method can be applied to separate their amplitudes and quantify their volumes. However, in fines and clays, where pores are smaller, the water protons relax faster due to the surface relaxation at the water–rock interface, thus generating the signal in the fast-relaxing part of distribution where it can overlap with the signal originating from heavy oil and bitumen (Fig. 1B). In addition

to that, the diffusion coupling effect may further decrease the interpretability of the oil and water signals. This effect occurs in saturated and connected micro- and macropores when water is in diffusional exchange, causing the change in the relationship between $T_2$ relaxation and pore size distribution[14]. In strong diffusive-coupling conditions, macro- and micro pore water signals will merge into a single peak, rendering the single $T_2$ cutoff and deconvolution approach inaccurate[15].

In these cases, signal deconvolution can be challenging, requiring the application of a more sophisticated separation method.

Among other factors influencing the $T_2$ distribution of saturated core samples, the reports in the literature showed that the relative increase or decrease of fluid saturation produces systematic shifting in sample $T_2$ distribution which can be modelled and used for monitoring water saturation change[16]. As surface relaxation and diffusive coupling play a vital role in NMR petrophysical studies of porous media, the study findings also confirmed the change of dominating relaxation mechanisms with sample saturation. Another factor influencing water fraction quantification is the clay content and type[17]. Studies show that different types of clays such as illite, smectite, and kaolinite can be distinguished from NMR measurements, particularly by 2D mapping. As these clay minerals adsorb and bond variable amounts of water molecules, knowledge of their relative fraction could contribute to the more precise differentiation of bound and producible fluids from NMR measurements.

**XGBoost principles.** XGBoost stands for Extreme Gradient Boosting (XGB), and it presents an implementation of the gradient boosting decision trees[18]. The main principle of gradient boosting is to utilize the individual weak learner, such as decision tree, and in a stage-wise manner, add iteratively new trees, to minimize further the objective function. This process continues for the specified number of boosting iterations, after which the prediction model is obtained in a final form. To achieve this effectively, the algorithm uses gradient descent to minimize the objective function, by finding the direction of the "steepest" descent. XGBoost on the other hand employs a number of improvements resulting in better overall generalization and computational speed. Some of the key advantages are the use of second-order gradients which contribute to a better understanding of the direction of the loss function minimum, and enhanced regularization techniques such as lasso regression (L1) and ridge regression (L2) which reduce the model complexity and overfitting. The XGBoost model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

where $\hat{y_i}$ is predicted Dean-Stark water content (%DS-w), $x_i$ is a vector of input features and $f_k$ is a tree at the $k$-th instance. A new tree $f_t$ is added iteratively by minimizing the objective function as:

$$f_t = \sum_{i=1}^{n} L\left(y_i, y_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

where $L$ presents the specified loss function, $y_i$ is observed %DS-w in a sample, $y_i^{(t-1)} + f_t(x_i)$ is the predicted %DS-w at the t−1 iteration, and $\Omega$ is a regularization term, or a penalty function. Regularization term can be denoted as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

where $\gamma$ is L1 and $\lambda$ is L2 regularization parameters, $T$ is the number of leaf nodes in a tree, and $w_j$ are the weights of leaves. The addition of new trees $f_t$ is performed in a stage-wise manner such that the loss between the prediction and observation is minimized, with respect to the regularization term $\Omega$ to prevent the overfitting and gauge the model complexity. Smaller values of $\Omega$ enable the better generalization of a tree. The detailed mathematical description of the XGBoost algorithm and additional tuning and regularization parameters is available elsewhere[18].

## Methodology

**Experimental procedure and data preprocessing.** Oil-sand samples were collected in northern Alberta in Canada, from a single delineation well. Two sets of 82 whole core samples were recovered. The first set was used for laboratory LF-NMR measurements, and a second set represented sister samples used in Dean-Stark extraction for determining the relative fraction of water, oil, and solids. Samples for NMR experiments were stored in glass vials, and measured using a Corespec 1000™ benchtop LF-NMR relaxometer, at reservoir temperature (6 °C) and ambient temperature (25 °C). The Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence was used for obtaining $T_2$-relaxation distribution. The CPMG parameters were predetermined after a series of test NMR experiments on different oil-sand samples. There were two aspects which had to be taken into account. The first was to tune the CPMG parameters to detect the fast relaxing heavy oil and clay-bound water signals. This was achieved by setting the shortest echo time TE that the equipment allowed (0.2 ms). The second aspect was achieving a lower signal-to-noise ratio (SNR) to simulate the well-logging in-situ NMR tool output by reducing the number of trains, which in turn resulted in a noisier signal. After trial rounds of measurements, the following parameters were found to produce optimal $T_2$ distribution and SNR (Table 1).

For the dataset, the range of SNR varied from 5 to 56 with an average of 23. The ExpFit in-house software for multi-exponential analysis of the NMR signal was used. The representation of the signal after Inverse Laplace Transform (ILT) was obtained using Tikhonov regularization[19]. The practice has shown that the regularization parameter helps avoid oscillations in solution associated with noise and provides smooth $T_2$ distributions[20].

| CPMG pulse parameters | Values |
|---|---|
| Echo time, TE (ms) | 0.2 |
| Number of pulses, Np | 5000 |
| Wait time/post train delay (ms) | 6500 |
| Number of trains, Nt | 10 |

**Table 1.** Optimal CPMG pulse sequence parameters for detection of fast relaxing clay-bound water and heavy oil signals.

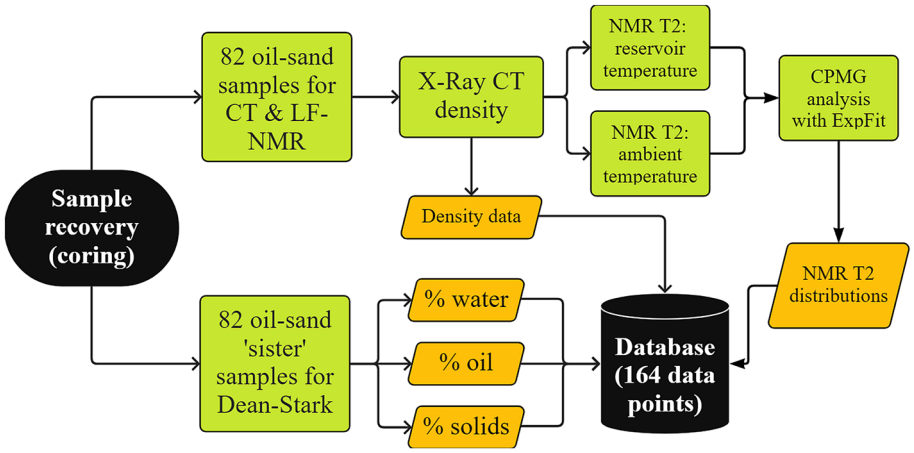


**Figure 2.** Flowchart representing the experimental program for oil-sands samples, by X-ray CT, LF-NMR $T_2$ measurements and Dean-Stark extraction.

The regularization parameter can be determined by direct and indirect methods such as Butler-Reed-Dawson, L-curve, or generalized cross-validation[20]. In the case of oil-sands, after initial analysis, the regularization parameter was determined directly and $\alpha = 0.05$ was found to provide the most stable solution for most of the samples. The density values of these samples were measured beforehand by X-ray Computed Tomography (X-ray CT) using GE 9800 CT scanner, as a substitute for the density logging data.

The experimental program for Dean-Stark extraction, LF-NMR measurements, and X-ray CT density measurements is illustrated in the flowchart (Fig. 2).

The final size of the dataset comprised 164 data points—82 $T_2$ distributions at ambient temperature, and 82 $T_2$ distributions at reservoir temperature, with corresponding density data and Dean-Stark sample composition. To compare the performance of machine learning models with the well-known peak deconvolution approach, the prediction of water content by LF-NMR measurements was also performed using the $T_2$ cutoff approach developed by Bryan et al.[7].

The data processing and model training was performed in Python 3.9 environment, while figures were produced using OriginPro 2019b software. For XGBoost model development and training, the dataset was randomly split with data shuffling into a training set and a test set in 0.25:0.75 proportion respectively. To ensure the reproducible split of the data, a random split seed was fixed to *random_state* = *2*. The XGBoost models were optimized using Bayesian Optimization (BO), while the training quality was evaluated by leave-one-out cross-validation (LOOCV). The forecasting performance of the models was evaluated using three error metrics, and residual distribution analysis. These steps will be discussed in detail in the following sections.

**XGBoost model based on feature engineering (XGB-FE).** Feature engineering (FE) is a process in a part of a machine learning pipeline where domain knowledge is utilized to extract the most relevant information from the raw data. In this work, we used feature engineering to extract information from the NMR $T_2$-relaxation distribution. The complete FE model derivation procedure is illustrated in Fig. 3.

In petrophysics, the $T_2$-relaxation is regularly analyzed by geoscientists to determine fluid saturations in reservoirs, differentiate between different types of fluids, study the distribution of pore size, and evaluate physicochemical properties of fluids. However, depending on the task, some parts of the $T_2$ distribution may have more relevance than others. In the context of studying the water content in oil-sands by NMR, we use feature engineering to reduce the amount of unnecessary information while preserving the data carrying the most information about the water in samples. A time-domain distribution of the $T_2$-relaxation was obtained by processing the spin-echo signal decay using a mathematical inversion. As the $T_2$ distribution has a form of a continuous function, the discretization was performed for data binning which simplifies the input of data into the machine learning model. After the discretization, the $T_2$ data was presented as a frequency distribution by 52 bins, with

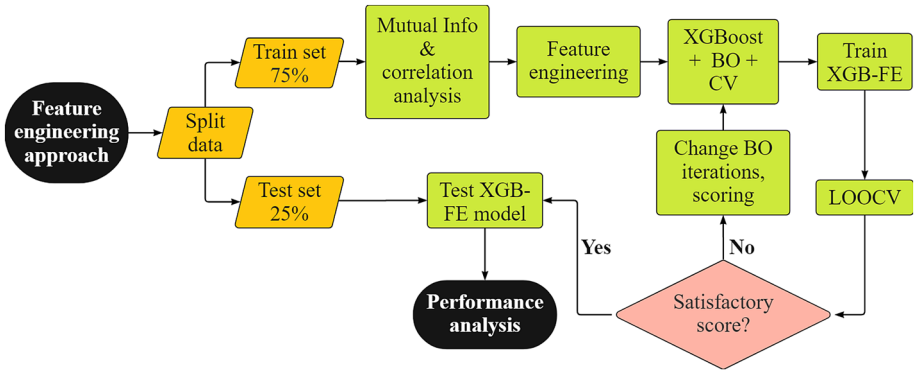


**Figure 3.** Flowchart for XGB-FE model development.

each bin corresponding to a particular $T_{2i}$ relaxation time in milliseconds. To limit the number of inputs, we defined five new NMR $T_2$ features.

As the $T_2$ distribution of relaxation times is represented on the semi-logarithmic scale, the standard parameter for representing the average $T_2$ relaxation is $T_2$ logarithmic mean ($T_{2lm}$):

$$T_{2lm} = \exp\left[\sum \frac{A_i}{A} \cdot \ln\left(T_{2i}\right)\right] \tag{6}$$

where $A_i$ is an amplitude at the corresponding $T_{2i}$ bin, and A is a total NMR amplitude. Empirical evidence shows the strong relationship between viscosity of fluids and $T_{2lm}$, implying that in a water–oil system where distribution tends to be multimodal due to their different relaxation properties, the $T_{2lm}$ provides a better measure central tendency favoring both fast and slow relaxing parts of the distribution.

To account for the variation in $T_2$ distribution (i.e. narrow vs. wide peaks), the $T_2$ standard deviation was defined as:

$$T_{2std} = \sqrt{\frac{\sum(A_i - \mu)}{N}} \tag{7}$$

where $\mu$ is the $T_2$ distribution mean, and $N$ is the number of the $T_2$ bins.

The $T_{2p}$ was defined as a location of a maximum value (peak) of the $T_2$ amplitude on $T_{2i}$ axis. This parameter is used in the petrophysical practice for the separation of bound and producible fluids and fluid typing, since $T_{2p}$ gives an indication of whether the largest amplitude portion of the signal corresponds to low or high $T_2$ values.

$$T_{2p} = \max\left(f\left(T_{2_1}\right), \ldots, f\left(T_{2_n}\right)\right) \tag{8}$$

Intelligent algorithms like XGBoost have gained popularity due to their ability to generalize complex data dependencies in large datasets and achieve state-of-the-art forecasting results. However, a small dataset is used in this study, where the overlapping of water and oil $T_2$ signals are likely to remain hidden or poorly represented. So, instead of allowing the algorithm to search through the whole NMR $T_2$ distribution, we can 'show' it where to look for the patterns and changes in the amplitude. One of the important parts of the $T_2$ distribution in sandstones is the empirical clay-bound water $T_2$ cutoff located at 3 ms, which presents the boundary between capillary bound and clay-bound fluids in a water-saturated core[21,22]. In order to capture the possible $T_2$ response of clay-bound water, and monitor its signal variation with different training samples, we defined a $T_2$ bound fluid ($T_{2bf}$) interval as:

$$T_{2bf} = \sum_{0.1(ms)}^{3.0(ms)} A_i \tag{9}$$

However, this parameter cannot be used on its own to describe the changes in water content, since the oil signal may also be located in the relevant interval. The true $T_2$ cutoff value in petrophysical practice is usually determined by performing lab tests on the saturated core samples (i.e. centrifuging), and even then the use of a fixed or averaged $T_2$ cutoff value leads to the erroneous prediction of producible fluids. Instead, we attempt to obtain insights about the true $T_2$ cutoffs using a feature extraction technique called Mutual Information (MI) regression, based on the information entropy between variables. In classical regression analysis, statistical tests like F-test are carried out to study the degree of the linear association or continuous analysis of covariance (CANOVA) for the non-linear association between variables. However, mutual information is not 'concerned' whether the variables have apparent linear correlation or covariance of zero, and they may still be stochastically dependent. This is the case in studying the changes in the conditional probability of one variable when another is modified[23]. In other words, by using MI regression, one can measure the level of association of the specific parts of $T_2$ distribution with the target output (i.e. water content by Dean-Stark), regardless of their correlation or covariance. The score is measured in natural units of information or 'nats' which are based on natural

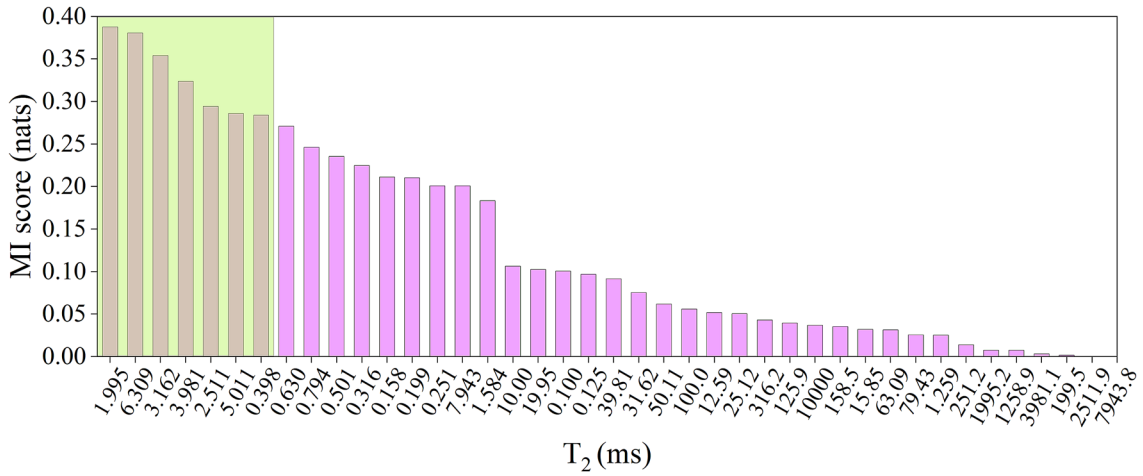


**Figure 4.** Results of the mutual information regression applied to the training set $T_2$ distributions of the oil-sand samples relative to the Dean-Stark water content (DS-w). The shaded area presents the continuous cluster of $T_2$ responses with a strong mutual association with DS-w, which were used for the calculation of the $T_2$ cutoff range parameter – $T_{2cr}$.

logarithms and powers of *e*. The MI regression was performed on the training set using a Python library sklearn. feature_selection class mutual_info_regression.

Figure 4 shows the relative mutual information scores of $T_2$ responses, where higher values indicate a stronger association with water content by Dean-Stark. For this dataset, the responses from 1.99 to 6.30 ms have the highest association with the water signal and form a continuous cluster between $10^0$ and $10^1$ decades along the $T_2$ semi-log scale, suggesting that most of the theoretical $T_2$ cutoff values lie in this interval. Therefore, the $T_2$ cutoff range ($T_{2cr}$), was defined as:

$$T_{2cr} = \sum_{1.99(ms)}^{6.30(ms)} A_i \tag{10}$$

As previously mentioned, the $T_2$ surface relaxation and diffusive-coupling play an important role in the identification of clay-bound water, which causes the overlapping of the water and oil signals. Unfortunately, for the determination of their contribution, a sample recovery for the subsequent lab experiments is required. However, a common practice in well-logging is to combine NMR and bulk density logs to improve interpretation. Therefore, the bulk density was used as an additional parameter which we postulate is associated with $T_2$ surface relaxation and diffusive-coupling.

The correlation matrix (Fig. 5) shows the amount of linear dependence between the input features and target output. According to the Pearson score, $T_2$ cutoff range and $T_2$ peak and $T_2$ logarithmic mean features exhibit the strongest positive correlation with the water content by Dean-Stark (DS-w). The $T_2$ standard deviation shows a moderate degree of positive correlation, while $T_2$ bound fluid and density features show moderate to low negative correlation with DS-w. Interestingly, when compared with mutual information scores from Fig. 5, it can be observed that all features are ranked by score accordingly to Pearson's scores except for density which has the highest MI score (0.86 nats), indicating its strong stochastic (nonlinear) dependence with DS-w, thus justifying integration of density measurements into the model. Therefore, the XGB-FE model was developed using the six features presented in Table 2 (Fig. 6).

**XGBoost model based on the full $T_2$ relaxation distribution (XGB-FS).** The second modeling approach facilitates the complete sample $T_2$ distribution. There are two main incentives for this approach. First, the $T_2$ relaxation distribution contains a large amount of information about the fluids residing in the pore space, indicating that the use of a single or even a few features to characterize the whole distribution, may lead to the significant information loss, and therefore to poor model forecasting performance[24]. By using the entire $T_2$ distribution, variations such as changes in slope or local minima can implicitly be used to help separate oil and water signals. Secondly, predictions generated by the full-$T_2$ distribution model provide a good baseline for comparison with the feature engineering approach and conventional deconvolution approach. Therefore, the input features were arranged as $X = [A_1, A_2, A_3, \ldots, A_{52}, \rho_i]$ where $A_i$ is the $i$-th column vector of the amplitudes at the corresponding $T_{2i}$ bin, and $\rho_i$ is a column vector of density measurements. The water content by Dean-Stark (DS-w) was arranged as $Y = [DS\text{-}w_1, DS\text{-}w_2, \ldots, DS\text{-}w_n]$ thus defining the dataset as $\{(X_i, Y_i)\}_{i=1}^{n}$ where $n$ is the number of oil-sands samples. The complete XGB-FS model derivation procedure is illustrated in Fig. 7.

**Model optimization.** The XGBoost algorithm contains many hyperparameters which enable fine model tuning. From the standpoint of statistical learning, the tuning usually involves the use of iterative algorithms which search for a suitable combination of hyperparameters in real-valued parameter space, relative to the speci-

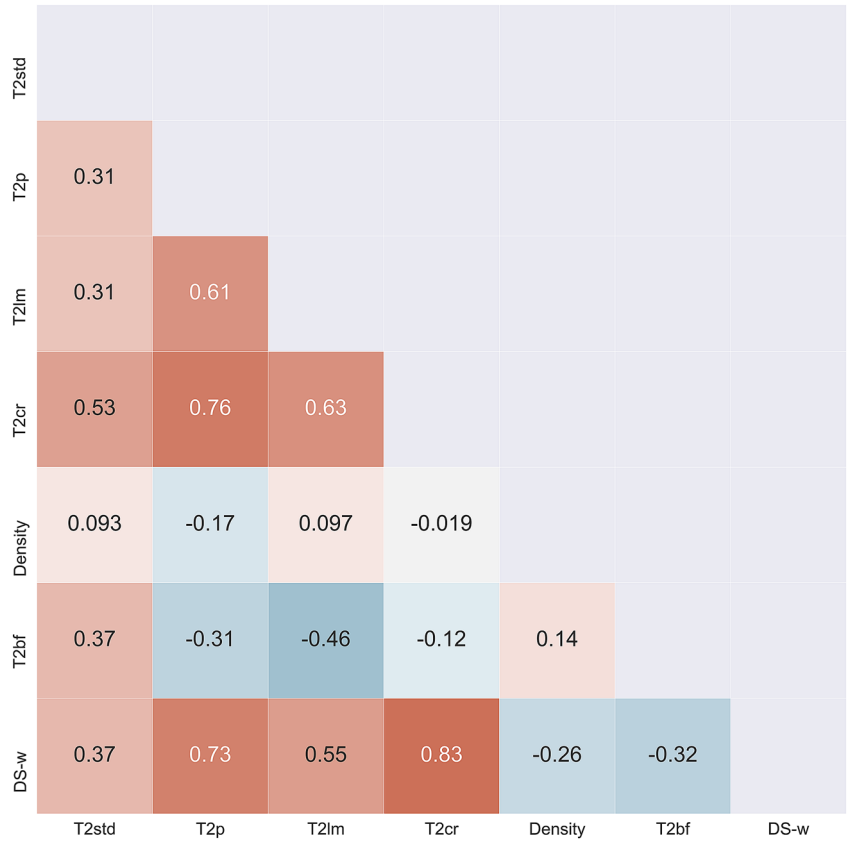


**Figure 5.** The diagonal correlation matrix showing the amount of linear dependence between six input features with Dean-Stark water content (DS-w). Scores represent the Pearson's correlation coefficient and are color coded (heatmap).

| Statistic | $T_{2std}$ (a.u.) | $T_{2p}$ (ms) | $T_{2lm}$ (ms) | $T_{2cr}$ (a.u.) | $T_{2bf}$ (a.u.) | ρ (kg/m$^3$) |
|---|---|---|---|---|---|---|
| Count | 164.000 | 164.00 | 164.00 | 164.00 | 164.00 | 164,00 |
| Mean | 0.016 | 11.57 | 1.84 | 0.18 | 0.31 | 1626 |
| Std | 0.005 | 4.54 | 1.22 | 0.13 | 0.10 | 80 |
| Min | 0.006 | 1.00 | 0.32 | 0.00 | 0.10 | 1442 |
| 25% | 0.013 | 8.00 | 0.90 | 0.07 | 0.23 | 1581 |
| 50% | 0.016 | 13.00 | 1.61 | 0.18 | 0.30 | 1634 |
| 75% | 0.019 | 15.25 | 2.49 | 0.29 | 0.38 | 1677 |
| Max | 0.032 | 20.00 | 8.59 | 0.50 | 0.54 | 1842 |

**Table 2.** Descriptive statistics of six input features used for XGB-FE model development.

fied measure of model forecasting performance (e.g., mean squared error). However, as the number of parameters grows, the optimization becomes computationally expensive due to the combinatorial explosion, making the manual optimization or exhaustive grid searching techniques inefficient. In contrast, Bayesian Optimization (BO) sets a probabilistic approach where each successive combination of hyperparameters is selected based on the information obtained in the previous optimization step, thus avoiding the redundant calculations for unlikely parameter combinations and reducing the number of required iterations to reach the global minimum of the objective function.

The BO was performed in Python using scikit-optimize package class skopt.BayesSearchCV. The hyperparameters and their optimal values are presented in Table 3.

**Performance metrics and model validation.** The forecasting performance of the models was evaluated using three performance metrics including coefficient of determination ($R^2$), root mean squared error (RMSE), and mean absolute error (MAE). The $R^2$ is the positively oriented metric used in regression for representing the

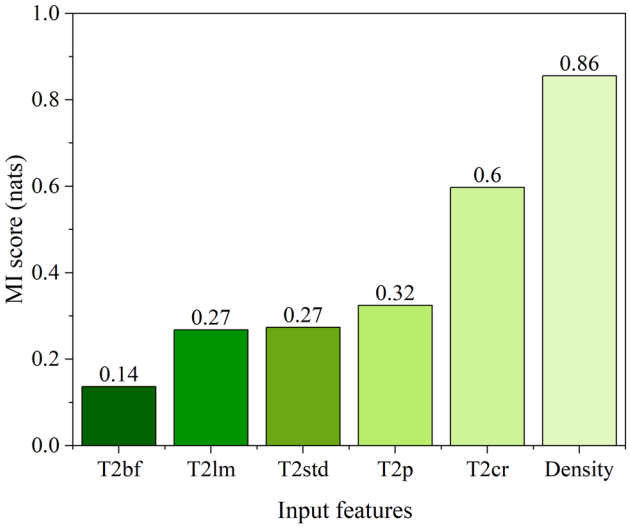


**Figure 6.** Mutual information regression scores for five NMR parameters and bulk density (input features) relative to the Dean-Stark water content (DS-w).

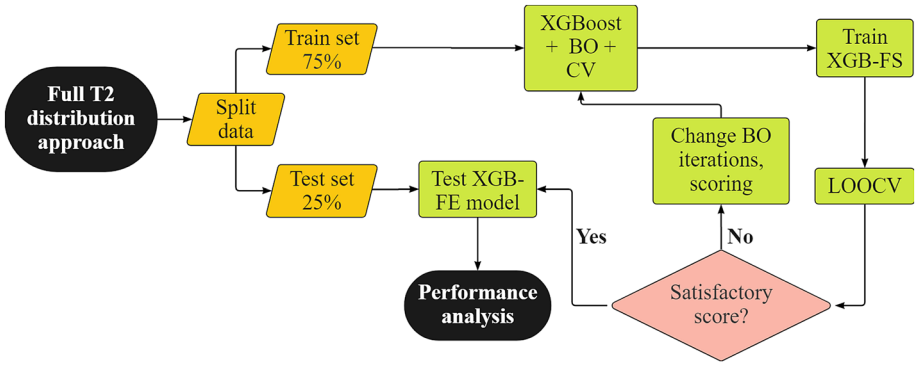


**Figure 7.** Flowchart for XGB-FS model development.

| XGBoost hyperparameters | Search range | XGB-FS optimal | XGB-FE optimal |
|---|---|---|---|
| n_estimators | [50–1000] | [650] | [300] |
| learning_rate | [0.004–0.1] | [0.008] | [0.053] |
| subsample | [0.7–1.0] | [0.7] | [0.6] |
| max_depth | [6–12] | [8] | [7] |
| objective | ['squared_error', 'pseudo_huber'] | ['pseudo_huber'] | ['squared_error'] |
| grow_policy | ['depthwise', 'lossguide'] | ['lossguide'] | ['lossguide'] |
| booster | ['gbtree', 'dart'] | ['gbtree'] | ['gbtree'] |

**Table 3.** Results of Bayesian Optimization with fivefold cross-validation, for XGB-FS and XGB-FE models.

amount of model variance, and how well the model predictions generalize the observations. However, $R^2$ alone does not provide information on prediction errors. The RMSE is another regularly employed error metric, used alongside $R^2$, but under the assumption that residuals follow the normal distribution[25]. As a result of the heavy penalization of larger residuals, the RMSE is a convenient metric for revealing the differences in performance between multiple models with normally distributed residuals. At the same time, large residuals can cause the inflation of the RMSE score, which is why MAE can be used for additional evaluation. The MAE measures the mean magnitude of model prediction errors, but in contrast to RMSE, the errors are not squared. Therefore RMSE scores are always greater or equal to MAE scores. These two metrics can be used together to estimate the variation in errors, where RMSE = MAE indicates no variation in the magnitude of prediction residuals. Note that both RMSE and MAE are negatively oriented scores expressed in %DS-w.

$$R_2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{13}$$

where $y_i$ is predicted %DS-w, $\hat{y}_i$ is observed %DS-w, $\bar{y}$ the sample mean, and $n$ presents the number of samples.

The further model performance evaluation and validation were performed using leave-one-out cross-validation (LOOCV) due to its convenience for use on small datasets (Fig. 10). Cross-validation is a resampling method in which the sample subsets are drawn repeatedly from the training set, followed by model refitting for each subset, thus providing information on model fitting variability. In LOOCV, the samples are drawn for one observation at a time, while the rest of the data is used for model training. Therefore, this process has a number of iterations equal to the number of samples, making it computationally expensive for large datasets.

In addition to LOOCV, the permutation tests were conducted for assessing the significance of fivefold cross-validated model prediction scores with 150 random permutations. This enabled the evaluation of the statistical significance of model predictions and their inputs by a permutation test *P*-value.

## Results

In this section, the performance of three models is presented including the XGB-FE model based on the XGBoost algorithm with feature engineering, the XGB-FS model based on the XGBoost algorithm using the whole sample $T_2$ distribution, and a peak deconvolution approach (Bryan et al[7]). To assess the model performance in more detail, residual plots (Fig. 82,5,8) and quantile–quantile plots (Fig. 83,6,9) are used for the analysis of the residual normality, and model variance and bias. All results are summarized in Figs. 8, 9 and 10.

Analysis and comparison of error statistics, cross-plots, and distribution of residuals indicate that the XGB-FE model achieves the highest accuracy and generalization ability in the prediction of water content in oil-sand samples. Figure 8-1 shows that apart from slight overprediction in the 3–5% DS-w range, all XGB-FE predictions spread along the x = y line, with low variance, achieving the highest $R^2$ score in the study ($R^2 = 0.90$). Figure 8-2 shows the constant low variance of the residuals, indicating that the model inputs capture variation in the data properly. Finally, the Q-Q plot (Fig. 8-3) confirms the residual normality and thereby the underlying assumption that XGB-FE model residuals follow the normal distribution (low bias, low variance). Finally, the XGB-FE model achieves 1.5–3 times lower RMSE and MAE scores compared to the XGB-FS and Bryan et al.[7] models indicating the best generalization ability of the three.

As for the XGB-FS model predictions, Figs. 8-4,5,9 show a similar residual distribution to XGB-FE (normality and bias). The residual variance however is increased but constant, therefore achieving a somewhat lower $R^2$ score ($R^2 = 0.85$), and 1.5 times higher RMSE and MAE compared to XGB-FE. From Fig. 8-7, it can be observed that the Bryan et al.[7] model generally tends to underpredict the water content in samples. In addition, Fig. 8-8,9 show inflated but constant variance in the distribution of residuals, while residual normality still holds with some local perturbing. As a result, Bryan et al.[7] model RMSE and MAE scores are the highest of the three (Fig. 9).

## Discussion

The two machine learning models in this study were designed to test two principal hypotheses. First, to confirm that integration of density measurements into the machine learning models can help to separate the contribution from overlapping oil and water signals. Second, to show that the derivation of new LF-NMR $T_2$ features can improve the generalization ability of the machine learning model to a degree that can enable the accurate forecasting of water content by LF-NMR in oil wells (in-situ).

Bulk density measurements are regularly used together with LF-NMR measurements in petrophysical practice to improve the interpretation of well logs[4,26]. LF-NMR measures the response of the fluids in the rock pore space, therefore carrying information about the fluids and pore size distribution of the rock. On the other hand, density logging equipment measures the response of the solids (rock matrix), together with fluids. The two are related in terms of $T_2$ surface relaxation which depends on the rock pore to surface ratio with the fluid volume, as well as with diffusive-coupling effect. This dependence can be also observed from the prediction test scores of the XGB-FS and XGB-FE models with and without bulk density as one of the model inputs. Prediction scores from Tables 4 and 5 indicate that models achieve better scores with the integration of bulk density, therefore confirming the relationship between Dean-Stark water content and density discovered by mutual information regression.

To affirm the second hypothesis: when XGB-FS and XGB-FE are compared (Fig. 8), it can be observed that XGB-FE achieves better performance, especially in terms of prediction variance. The XGB-FE variance reduction supports two premises. First, the engineered features properly capture all the relevant information from the $T_2$ distribution, indicating negligible information loss. Secondly, in the feature engineering case, the XGBoost algorithm generalized the variability in the data with the output more effectively, suggesting that for smaller datasets the appropriate feature engineering enables the XGBoost algorithm to discover dependencies within the data more effectively than for a large number of raw information (53 features in case of XGB-FS). In other words, the new features contain all the relevant parts of the $T_2$ distribution compressed into a few values, which

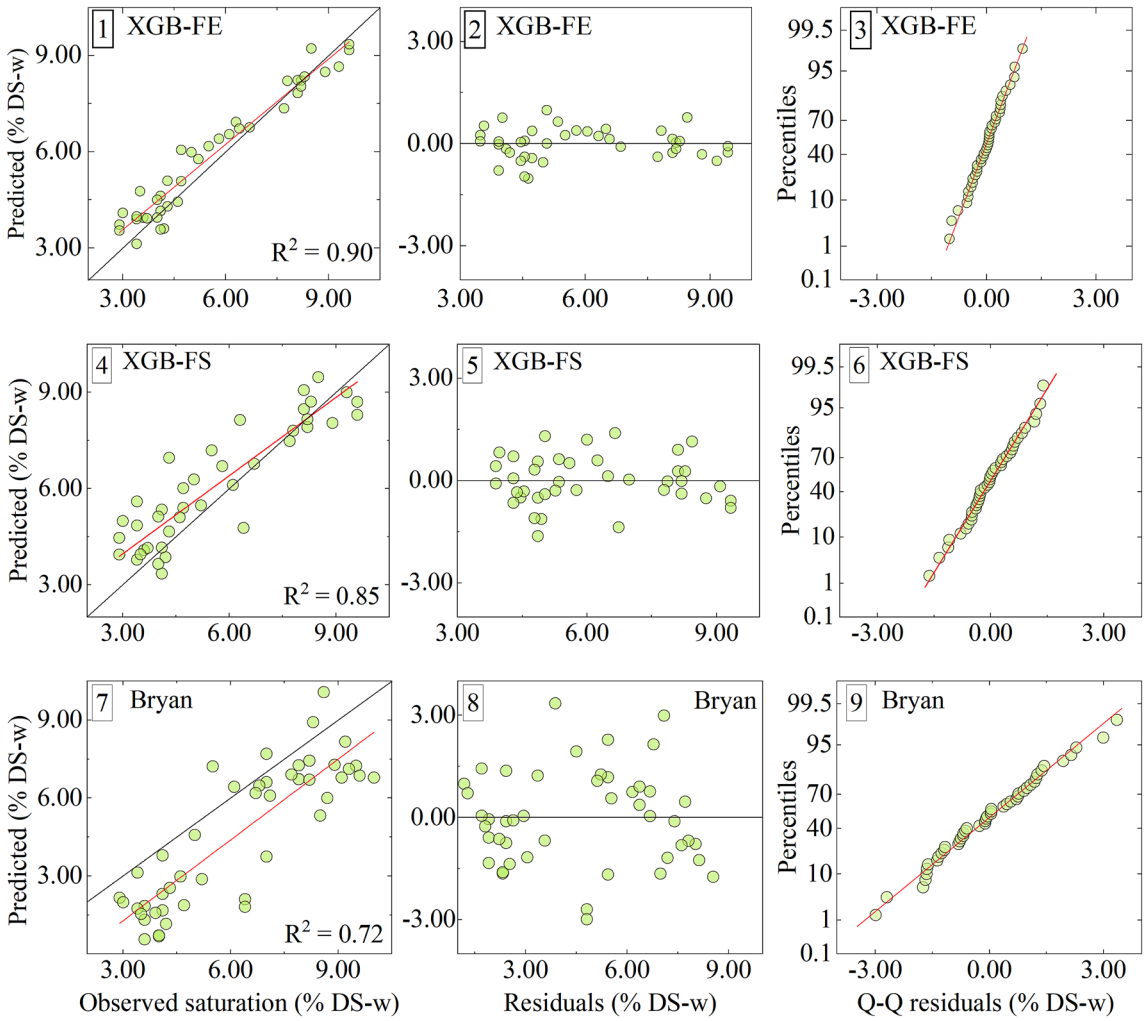


**Figure 8.** Evaluation of XGB-FE, XGB-FS and Bryan et al.[7], performance by cross-plots between the model predictions and observed saturation in %DS-w (**1**, **4**, **7**), distribution of regular residuals (**2**, **5**, **8**), and quantile–quantile plots for comparing distributions between predictions and observations and evaluating normality of residuals (**3**, **6**, **9**).

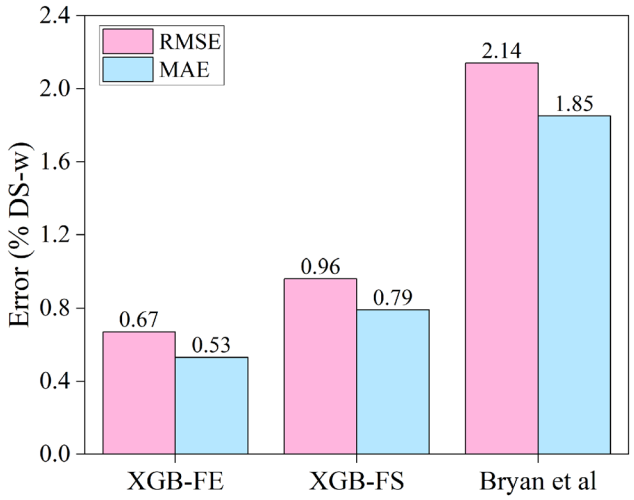


**Figure 9.** Comparison of RMSE and MAE test prediction scores for the three models (*'random_state = 2'*).

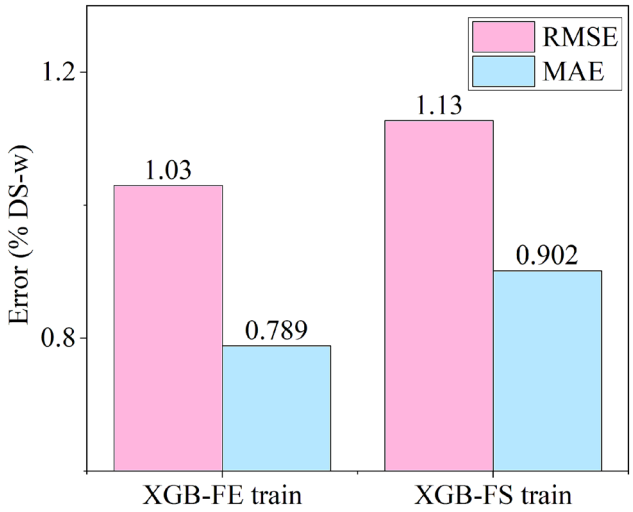


**Figure 10.** Leave-one-out cross-validation (LOOCV) scores for XGB-FS and XGB-FE machine learning models for the training set with fixed random split seed *'random_state = 2'*. Note y-axis was truncated for convenience.

| XGB-FS | | |
|---|---|---|
| **Statistic** | **wo/Density** | **w/Density** |
| RMSE (%DS-w) | 1.08 | 0.96 |
| MAE (%DS-w) | 0.86 | 0.79 |
| $R^2$ | 0.72 | 0.85 |

**Table 4.** Comparison of XGB-FS model performance with and without *bulk density* parameter.

| XGB-FE | | |
|---|---|---|
| **Statistic** | **wo/Density** | **w/Density** |
| RMSE (%DS-w) | 0.91 | 0.67 |
| MAE (%DS-w) | 0.73 | 0.53 |
| $R^2$ | 0.81 | 0.90 |

**Table 5.** Comparison of XGB-FE model performance with and without *bulk density* parameter.

ultimately reduces the XGB-FE model complexity and therefore enables better generalization of the relationship between inputs and a target variable (DS-Sw). According to Fig. 6, along with a bulk density (MI = 0.86 nats), the $T_2$ cutoff range feature ranks second by MI score (0.60 nats), indicating the variability of the sum of $T_2$ responses between 1.99 and 6.30 ms has a strong relationship with water signal. The location of the $T_2$ peak ($T_{2p}$), $T_2$ standard deviation of the spectrum ($T_{2std}$), and $T_2$ logarithmic mean ($T_{2lm}$) achieve similar MI scores (0.27 and 0.30 nats, respectively) signifying that these features alone do not capture enough information about the water content. Finally, the sum of $T_2$ responses representative of the empirical clay-bound water part of the $T_2$ distribution (0.1–3.0 ms) shows the least association with the target (DS-w). Although these features alone cannot explain variance in data effectively, their mutual interaction can improve it. Since MI does not consider this mutual interaction between features relative to the target output, the correlation matrix can be used. For instance, Fig. 3 shows that the $T_{2cr}$ versus $T_{2p}$ and $T_{2cr}$ versus $T_{2lm}$ have a strong positive correlation (0.76 and 0.63 respectively), while $T_{2bf}$ versus$T_{2lm}$ have a moderate negative correlation. These interactions are likely to be generalized in the XGB-FE model training process, thus explaining its improved performance. Furthermore, the permutation test score of XGB-FE using 150 permutations generated a *P*-value of 0.001, compared to the XGB-FS P-value of 0.007. In both cases the *P*-value is well below 0.05, showing a very low likelihood of obtaining such model performance purely by chance.

As for the deconvolution approach (Bryan et al.[7]), the main challenge lies in the separation of overlapping fluid contributions in $T_2$ distribution. Even under the assumption that $T_2$ cutoff and deconvolution are performed such that a precise distinction between fluid signals is possible, the issue of how to associate the amplitudes with respect to mass persists. For the given dataset, this approach leads to underprediction of water content, indicating that the oil and water signals are not sufficiently separated. The machine learning-based approach is more robust

because it removes the necessity to manually identify peak separation and the errors associated with visually separating oil and water signals, especially in the case of NMR measurements acquired at low SNR.

It is important to point out the limitations of these models which are related to reservoir lithology (1) and SNR of the measurements (2):

(1) The models presented in this study were derived for the oil-sands reservoir, which is why their application is limited only to similar reservoir types. However, the presented approach can be extended for use in other types of oil reservoirs, under the assumption that a sufficiently large amount of observations is available.
(2) The SNR achieved by the benchtop LF-NMR relaxometers can be up to 30 times higher than the SNR values obtained using well-logging tools. In this study, the NMR signal-to-noise ratio was on average 20, which can be still considered high relative to the logging tools where the SNR of 3–5 is considered satisfactory[27]. Although the recent research demonstrated that XGBoost algorithm is sufficiently robust even with noisy data[28], an additional validation using the data obtained by the LF-NMR logging tools would be desirable. It is worth noting that, in lower SNR samples, the deconvolution approach will be even more challenging, and the value of using just the general properties of the $T_2$ distribution and XGBoost may be even further enhanced.

As a follow-up study, the procedures for NMR measurements with a controlled saturation and desaturation of samples, similar to those reported in recent literature[10], would enable deeper sensitivity analysis of the features derived in this work and further improvement of the XGB-FE model. In such a setup, the Dean-Stark measurements could be replaced by the more cost and time-effective mass-volume measurements, ultimately allowing the collection of a larger database at which point the application of artificial neural networks (ANNs) would be possible.

It is also worth noting that logging equipment configuration can be substantially different from desktop NMR relaxometers, which may cause inconsistencies between NMR $T_2$ distributions obtained in the lab and the field. This can cause the variable performance of proposed NMR data-driven model, which is why the parameters of the NMR logging device, such as TW, TE and number of trains, should be relatively consistent to the values reported in this study.

## Conclusions

This study presents the approach which integrates extreme gradient boosting with LF-NMR measurements and bulk density data for the water saturation determination in oil-sands. Two models were developed using full NMR $T_2$ distribution (XGB-FS), and feature engineering (XGB-FE). It is concluded that;

- Feature engineering can be effectively used to extract vital information from NMR $T_2$ distribution, using domain knowledge and mutual information regression.
- The integration of bulk density data as a model input notably improves the XGB-FS and XGB-FE forecasting performance.
- XGB-FE achieved RMSE = 0.67%, MAE = 0.53% and $R^2 = 0.90$ in predicting relative water content by Dean-Stark, a substantial improvement compared to deconvolution method.

These results suggest that the XGB-FE model can be extended for the improved *in-situ* water saturation determination.

## Data availibility

Correspondence and requests for materials should be addressed to S.M.

Received: 29 March 2022; Accepted: 2 August 2022
Published online: 17 August 2022

## References

1. Donaldson, E. C. Well logging for earth scientists. *J. Pet. Sci. Eng.* **2**, (1989).
2. Alboudwarej, H. *et al.* Highlighting heavy oil. *Oilf. Rev.* **18**, 34–53 (2006).
3. Liu, J., Feng, X. Y. & Wang, D. S. Determination of water content in crude oil emulsion by LF-NMR CPMG sequence. *Pet. Sci. Technol.* **37**, 1123–1135 (2019).
4. Chen, J. & Bryan, J. In situ bitumen viscosity and saturation estimation from core log integration for Canadian oil sands. *Soc Pet. Eng. SPE Heavy Oil Conf. Canada* **3**, 1686–1693 (2013).
5. Venkataramanan, L. *et al.* An unsupervised learning algorithm to compute fluid volumes from NMR T1–T2 logs in unconventional reservoirs. *Petrophysics* **59**, 617–632 (2018).
6. Bryan, J., Kantzas, A., Bellehumeur, C. SPE 77329 viscosity predictions for crude oils and crude oil emulsions using low field NMR. *SPE J.* (2002).
7. Bryan, J., Mai, A., Hum, F. M. & Kantzas, A. Oil- and water-content measurements in bitumen ore and froth samples using low-field NMR. *SPE Reserv. Eval. Eng.* **9**, 654–663 (2006).
8. Mukhametdinova, A., Habina-Skrzyniarz, I., Kazak, A. & Krzyżak, A. NMR relaxometry interpretation of source rock liquid saturation — A holistic approach. *Mar. Pet. Geol.* **132**, 105165 (2021).
9. Newgord, C., Tandon, S. & Heidari, Z. Simultaneous assessment of wettability and water saturation using 2D NMR measurements. *Fuel* **270**, 117431 (2020).
10. Krzyżak, A. T., Habina-Skrzyniarz, I., Machowski, G. & Mazur, W. Overcoming the barriers to the exploration of nanoporous shales porosity. *Microporous Mesoporous Mater.* **298**, 110003 (2020).

11. Bai, Z. *et al.* Log interpretation method of resistivity low-contrast oil pays in Chang 8 tight sandstone of Huanxian area, Ordos Basin by support vector machine. *Sci. Rep.* **12**, 1046 (2022).
12. Ibrahim, A. F., Gowida, A., Ali, A. & Elkatatny, S. Machine learning application to predict in-situ stresses from logging data. *Sci. Rep.* **11**, 23445 (2021).
13. Li, H. & Misra, S. Long short-term memory and variational autoencoder with convolutional neural networks for generating NMR T2 DIstributions. *IEEE Geosci. Remote Sens. Lett.* **16**, 192–195 (2019).
14. Anand, V. & Hirasaki, G. J. Diffusional coupling between micro and macroporosity for NMR relaxation in sandstones and grainstones. *SPWLA Annu. Logging Symp.* **2005**(48), 289–307 (2005).
15. Singer, P. M., Chen, Z., Wang, X. & Hirasaki, G. J. Diffusive coupling in heptane-saturated kerogen isolates evidenced by NMR T1–T2 and T2–T2 maps. *Fuel* **280**, 118626 (2020).
16. Krzyzak, A. T. & Habina, I. Low field 1H NMR characterization of mesoporous silica MCM-41 and SBA-15 filled with different amount of water. *Microporous Mesoporous Mater.* **231**, 230–239 (2016).
17. Habina, I., Radzik, N., Topór, T. & Krzyżak, A. T. Insight into oil and gas-shales compounds signatures in low field 1H NMR and its application in porosity evaluation. *Microporous Mesoporous Mater.* **252**, 37–49 (2017).
18. Chen, T., Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 13-17-Augu, 785–794 (2016).
19. Tikhonov, A. N., Arsenin, V. Y. *Solutions of Ill-Posed Problems*. (V. H. Winston & Sons, 1977).
20. Testamanti, M. N. & Rezaee, R. Considerations for the acquisition and inversion of NMR T2 data in shales. *J. Pet. Sci. Eng.* **174**, 177–188 (2019).
21. Coates, G. R., Xiao, L., Prammer, M. G. NMR logging. *Ebooks* **253** (1999).
22. Prammer, M. G., Drack, E. D., Bouton, J. C. & Gardner, J. S. Measurements of clay-bound water and total porosity by magnetic resonance logging. *Log Anal.* **37**, 61–69 (1996).
23. Ross, B. C. Mutual information between discrete and continuous data sets. *PLoS ONE* **9**, e87357 (2014).
24. Ahmad, K. *et al.* Radial-basis-function-based nuclear magnetic resonance heavy oil viscosity prediction model for a Kuwait viscous oil field. *Interpretation* **4**, SF81–SF92 (2016).
25. Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **7**, 1247–1250 (2014).
26. Sun, B., Dunn, K. J., Latorraca, G. A., Liu, C. & Menard, G. Apparent hydrogen index and its correlation with heavy oil viscosity. *Annu. Logging Symp.* **298**, 1–14 (2007).
27. Jin, G., Xie, R., Liu, M. & Guo, J. Petrophysical parameter calculation based on NMR echo data in tight sandstone. *IEEE Trans. Geosci. Remote Sens.* **57**, 5618–5625 (2019).
28. Gómez-Ríos, A., Luengo, J., Herrera, F. A study on the noise label influence in boosting algorithms: Adaboost, GBM and XGBoost. In *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* Vol. 10334 LNCS, 268–280 (2017).


## Acknowledgements

This work was supported by the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-10-2020-119 within the development framework for a world-class Research Center. The authors want to gratefully acknowledge the support from the Center for Petroleum Science and Engineering (CPSE) from Skolkovo Institute of Science and Technology, WA School of Mines from Curtin University, and the Fundamentals of Unconventional Resources (FUR) group from the University of Calgary.

## Author contributions

S.M. and J.L.B. conceived the idea and collected the required data and participated in the methodology design. A.K. and J.L.B provided the samples and organized experimental work. S.M. performed data analysis, code writing, model optimization, and interpretation of the results. R.R, A.C., and A.T. also participated in results interpretation and provided supervision. D.K. performed machine learning framework validation. S.M. took the lead in writing the manuscript. All authors participated in revision and editing.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.M.