Data-Driven Workflow for the Preemptive Detection of Excess Water Producing Wells Drilled in Unconventional Shales

Yusuf Falola[1], Siddharth Misra[2], Jonathan Foster[3], and Mukul Bhatia[3]

[1]Harold Vance Department of Petroleum Engineering
[2]Harold Vance Department of Petroleum Engineering, Texas A&M University, USA
[3]The Department of Geology and Geophysics, Texas A&M University, USA

November 23, 2022

Abstract

The continuous rise in global energy demand requires the production of oil and gas from unconventional shale resources. One major concern has been the large volumes of produced water associated with the production of hydrocarbon from the shale resources. We developed a data-driven workflow for identifying potentially high water-producing wells drilled in unconventional shale formation. To that end, we applied unsupervised learning followed by supervised learning to process five conventional well logs, namely shallow and deep resistivity logs, density porosity logs, neutron porosity logs, and gamma ray logs, from a well drilled in an unconventional shale formation. A novelty of our study is the use of clustering methods to generate pseudo-lithology that is fed into a classifier for the desired identification of the excess water producing wells. The data-driven workflow was tested on 23 wells in Gulf coast basin and 29 wells in Fort Worth basin. Fort Worth and Gulf Coast basins in the U.S. are highly productive shale basins that produce 380 million cubic feet of gas and 1.74 million barrels of crude oil every day. Additionally, we identified geophysical signatures that explain the excess water production from the wells drilled in unconventional shale reservoirs. For future work, molecular simulation, core analysis, and advanced well logs studies need to be incorporated for a better explanation of the causes of excess water production in unconventional reservoirs.

Data-Driven Workflow for the Preemptive Detection of Excess Water Producing Wells Drilled in Unconventional Shales

Jonathan Foster, Siddharth Misra^*, Yusuf Falola^, Mukul Bhatia*[+]*

^Harold Vance Department of Petroleum Engineering, Texas A&M University, USA

*The Department of Geology and Geophysics, Texas A&M University, USA

[+]Berg-Hughes Center for Petroleum & Sedimentary Systems, Texas A&M University, USA

Corresponding author: misra@tamu.edu

ABSTRACT

The continuous rise in global energy demand requires the production of oil and gas from unconventional shale resources. One major concern has been the large volumes of produced water associated with the production of hydrocarbon from the shale resources. We developed a data-driven workflow for identifying potentially high water-producing wells drilled in unconventional shale formation. To that end, we applied unsupervised learning followed by supervised learning to process five conventional well logs, namely shallow and deep resistivity logs, density porosity logs, neutron porosity logs, and gamma ray logs, from a well drilled in an unconventional shale formation. A novelty of our study is the use of clustering methods to generate pseudo-lithology that is fed into a classifier for the desired identification of the excess water producing wells. The data-driven workflow was tested on 23 wells in Gulf coast basin and 29 wells in Fort Worth basin. Fort Worth and Gulf Coast basins in the U.S. are highly productive shale basins that produce 380 million cubic feet of gas and 1.74 million barrels of crude oil every day. Additionally, we identified geophysical signatures that explain the excess water production from the wells drilled in unconventional shale reservoirs. For future work, molecular simulation, core analysis, and advanced well logs studies need to be incorporated for a better explanation of the causes of excess water production in unconventional reservoirs.

I INTRODUCTION

Oil and gas production from unconventional reservoirs is essential for the global energy demand. Unconventional reservoirs comprise tight, low-permeability rocks, which restricts economical hydrocarbon production. The advancement in technology such as horizontal drilling and hydraulic fracturing have made such hydrocarbon deposits economically and operationally viable [1]. The US Energy Information Administration (EIA) estimates the total technically recoverable shale reserves to be around 420 billion barrels [2]. The United States accounts for 78.2 billion barrels of these reserves [3]. Most of the reserves and production are present in Marcellus, Barnett, Haynesville-Bossier, Fayetteville, Woodford, Antrim, Eagle Ford, Gulf Coast, and New Albany Shales [4].

The Fort Worth (FW) Basin in north-central Texas is a foreland basin and it extends to the southwestern corner of Oklahoma [5]. The Mississippian Barnett Shale which averages 4 wt.% total organic carbon (TOC) is the primary source rock for oil and gas in the basin [6,7]. It is one of the most productive shale plays in the US [8]. The United States Geologic Survey estimated that the Barnett shale contains 26.7 Tcf of gas, 1.1 billion barrels of natural gas liquids, and 98.5 million barrels of oil [5]. The Newark East field, discovered in 1981, produced virtually all the gas and condensate field from the Barnett shale, and it is considered the largest gas field in Texas [9]. The second most productive play in the Gulf Coast (GC) basin is the Late Cretaceous Eagle Ford group in the Texas portion [10]. The majority of the production comes from the Lower Eagle Ford Group (LEFG) [11]. Along, with the production of oil and gas, unconventional wells in the USA produce a lot of water. The volume of water produced from unconventional wells tends to be three times the volume of oil production [12].

Flowback (water associated with hydraulic fracturing) and produced (FP) waters are one of the main environmental and economic challenges faced in developing an unconventional reservoir. More than 90% of the FP waters are said be produced waters – naturally occurring formation brines extracted along with hydrocarbon [13]. These produced waters are shown to be allochthonous, they migrated into the shale formation before the conversion from smectite to illite [14]. Migration is possible due to presence of minor faulting and natural fractures beneath or above the hydrocarbon formation [15]. Disposal of the FP waters have been problematic, in many cases they are injected back into the subsurface which has been linked to increased seismic activity [16]. Additionally, FP waters contain organic compounds such as Benzine and BTEX which make unsafe for drinking and irrigation purposes [17]. Reuse of FP waters – for hydraulic fracturing for instance – has been suggested by researchers [18]; however, the logistics to handle the transportation of water from site to site due to uncertainty in volume of water to be produced remains a bottleneck [19]. The unit cost of handling the FP waters is expected to rise to over $5.00/bbl, which could make about 20% of unproduced barrels of oil in the Permian become uncommercial.

Machine learning, both supervised and unsupervised techniques, has been applied in tackling subsurface engineering problems related to hydrocarbon exploration and production. Supervised methods such as convolutional neural network (CNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression, and tree based methods have been used for reservoir properties prediction, facies classification, and history matching [20]. In addition, unsupervised methods such as K-means clustering, hierarchical clustering, spectral clustering, principal component analysis, and K-nearest neighbors have been used for facies classification, missing log data prediction, fracture detection from log data, and brittleness index estimation using well logs [21].

In this study, we leverage machine learning to process five conventional well logs – shallow and deep resistivity logs, density porosity logs, neutron porosity logs, and gamma ray logs – to establish a workflow for identifying potentially high water-producing wells in unconventional reservoirs. We improve a previously published work by incorporating clustering methods to generate pseudo-lithology that is fed into a classifier for the desired identification of the excess water producing wells [22]. The workflow uses unsupervised learning prior to supervised learning to accomplish the objective. Notably, in this study, we identified geophysical signatures responsible for the cause of excess water production from the wells drilled in unconventional shale reservoirs.

II THEORY AND METHODS

A DATA GATHERING AND PREPARATION

Five conventional well logs, namely shallow and deep resistivity (ILS and ILD, respectively), density porosity (DPHI), neutron porosity (NPHI), and gamma ray (GR) logs, were retrieved from 23 wells in the Gulf Coast basin and 29 wells in the Fort Worth basin. These five logs were used to develop and validate the proposed data-driven workflow. A challenging task in study was to find sufficient number of wells where all these 5 selected logs were acquired at a high quality. Notably, the wells are primarily from the Texas portion of these basins. Using the Kick Off Point (KOP) as a reference, logs from 200ft above the KOP and 300ft below the KOP were used in the data-driven workflow (see figure A1 in the Appendix A). The well log readings were recorded every 0.5ft interval along the wellbore. Furthermore, the production data from the last 2 years of the wells were used as targets to assist in developing the supervised models to identify the wells as High Water Producers (HWPs) or Low Water Producers (LWPs). Water Production Ratio (WPR) which is the ratio of produced water to total production was used as a distinguish factor between the HWPs and LWPs. A well with WPR higher than 0.7 is categorized as HWP, whereas a well with WPR lower than 0.5 is classified as LWP. It is not uncommon for well log data to contain outliers; thus, an outlier detection algorithm was employed to remove abnormal data points. In addition, most machine learning techniques based on distance and density calculations, such as K-means and K-Nearest neighbors, require feature scaling. So, the well logs were standardized to have a mean of 0 and unit variance. Resistivity logs were logarithmically transformed prior to clustering as a part of scaling.

B MACHINE LEARNING TECHNIQUES

Both supervised and unsupervised learning techniques were used in this study. First pseudo-lithologies were identified using unsupervised learning, next HWP/LWP wells were detected using supervised learning. Unsupervised methods – Kmeans and Agglomerative clustering – were used to generate pseudo-lithologies. A depth point in a well is assigned a pseudo-lithology using unsupervised learning. Next, supervised techniques, such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression, were used to simultaneously process the 5 logs and pseudo-lithologies of the depth points within the 500-ft depth interval around the kick-off point of a well to identify whether the well is high water producer (HWP) or low water producer (LWP). A well is assigned a label HWP or LWP using supervised learning. In summary, unsupervised followed by supervised learning is implemented in this study. A detailed description of the data-driven workflow employed in this study is presented in figure A2 in the Appendix A.

C PERFORMANCE METRICS

A key challenge in clustering (done prior to classification) is to determine the optimal number of clusters that signify the robustness and reliability of the unsupervised method in generating the pseudo-lithologies. To the end, silhouette, Calinski-Harabasz (CH), and Davies-Bouldin (DB) scores were used to determine the optimal number of clusters. Silhouette score ranges from -1 to 1, where 1 represents the best clustering performance. Calinski-Harabasz score has no range, but the higher the score, the better the clustering performance. In contrast, for Davies-Bouldin score, a value closer to zero represents a better clustering performance. In addition, the clusters generated by KMeans were compared with those generated using agglomerative clustering to determine the robustness of the clusters, i.e. pseudo-lithologies. For the comparison, adjusted rand score and homogeneity score were used, where values close to 1 indicate that different clustering methods based on distinct mathematical/statistical principles

are predicting the same clusters, which confirms the robustness of the clusters. After the evaluation of the clustering results, Matthews correlation coefficient (MCC) and F-1 scores were employed to determine the accuracy of the supervised learning models on the training dataset.

D FEATURE EXTRACTION AND REDUCTION

Features used for generating the pseudo-lithology of each depth using unsupervised learning are different from the features used for assigning the label (HWP or LWP) to a well. Features for the unsupervised learning were extracted from the individual and arithmetic combination of raw log data from the selected wells in the two basins (23 wells in Gulf coast basin and 29 wells in Fort Worth basin). These features were used for unsupervised learning to identify the pseudo-lithologies intersected by each well. The features used are listed in table 1. Subsequently, for the supervised learning, features were extracted from the depth and well log data by estimating statistical parameters such as mean, median, variance, kurtosis, and so on. Additionally, the presence of a cluster/pseudo-lithology and a count of more than 30 samples per cluster in a particular well were used as binary categorical features. For purposes of supervised learning, 393 features were generated for the GC basin while 337 features were generated for the FW basin. These features were reduced using univariate and bivariate statistical tests, such as Mutual Information classifier (MI) and analysis of variance (ANOVA) F-test. Such feature elimination or dimensionality reduction is essential to lower the variance and improve the generalization of the data-driven workflow. Figure A3 in the Appendix A summarizes the feature extraction and reduction for the supervised learning method.

TABLE 1: Features used for the first level and second level clustering of the two basins. Two-level clustering ensures that the dominant pseudo-lithologies don't bias the clustering results.

Region	**Gulf Coast**		**Fort Worth**	
Level	First Level Clustering	Second Level Clustering	First Level Clustering	Second Level Clustering
Clusters generated	A and B	A0, A1, A2, B0, and B1	A, B, C, and D	A0, A1, A2, B0, B1, and B2
Features used	DPHI GR Log10(ILD) (NPHI+DPHI)/2 Log10(ILS)×NPHI Log10(ILD)×DPHI	**Cluster A** DPHI GR Log10(ILD)×Log10(ILS) GR×DPHI Log10(ILS)-Log10(ILD) NPHI-DPHI **Cluster B** (NPHI+DPHI)/2 DPHI GR Log10(ILD) GR×NPHI	GR Log10(ILD) GR/DPHI Log10(ILD)×Log10(ILS) GR×NPHI	**Cluster A** Log10(ILD) ×Log10(ILS) DPHI Log10(ILS) Log10(ILS)-Log10(ILD) NPHI-DPHI GR×DPHI **Cluster B** DHPI GR NPHI GR×Log10(ILS) Log10(ILD)

III Results: Data-Driven Detection of High or Low Water Producers

For both the basins, a two-level clustering approach was adopted to ensure that the dominant pseudo-lithologies don't bias the clustering results. A dominant large-sized, high-density cluster adversely affects the quality of small-sized, low-density clusters. In other words, two-level clustering can reliably find the small-sized clusters without adverse effects due to the presence of large-sized, high-density clusters. In two-level clustering, the clusters obtained by the first application of clustering are further subdivided by applying the clustering method on the individual first-level clusters. The features used for the two levels

of clustering are listed in table 1. Best clustering performance is achieved when the features are customized for each cluster at each level. Table 1 lays out the various features used at various levels of clustering for different regions. The difference between neutron porosity and density porosity as well as the difference between logarithmic transformations of deep and shallow resistivity are important for sub-dividing the cluster A. Average of neutron porosity and density porosity is an important feature for generating the pseudo-lithologies for the Gulf-Coast region. Variations in features used at various levels and clusters ensure reliable and robust clustering.

A GENERATION OF PSEUDO-LITHOLOGIES USING UNSUPERVISED LEARNING

The results from the unsupervised methods are summarized in tables 2 and 3. Seven pseudo-lithologies were identified in the Gulf Coast basin, while six pseudo-lithologies were identified in Fort Worth basin. The overall silhouette scores for the two-level clustering are around 0.5. Silhouette scores indicate a decent clustering performance. The adjusted rand and homogeneity scores, which compare the clusters obtained using different clustering techniques, indicate excellent agreement. Both scores are close to 1, which confirms that the independent clustering techniques are predicting similar clusters. In summary, clustering results are robust and reliable.

TABLE 2: Reliability and robustness of the first level of clustering

	Clusters	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score	Adjusted Rand Score	Homogeneity Score
Gulf Coast	A, B, C, D	0.44	12281	0.90	0.99	0.98
Fort Worth	A, B	0.50	21603	0.53	0.99	0.97

TABLE 3: Reliability and robustness of the second level of clustering

	Cluster	Sub-Clusters	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score	Adjusted Rand Score	Homogeneity Score
Gulf Coast	A	A0, A1, A2	0.57	26666	0.56	0.91	0.87
	B	B0, B1	0.74	20885	0.46	1	1
Fort Worth	A	A0, A1, A2	0.57	54626	0.52	0.89	0.80
	B	B0, B1, B2	0.47	20395	0.75	0.90	0.85

For the first-level clustering, 4 clusters were obtained for the Gulf-Coast region and 2 clusters for the Fort Worth region. A comparison of silhouette, CH and DB scores between the two regions indicate a better performance for the Fort Worth region, which is obvious because only two clusters are obtained for the Fort-Worth region. For the second-level clustering, silhouette scores are higher than the first-level clustering, which confirms the need of the second-level clustering. A significant improvement in clustering is achieved for both the regions when the cluster B0 is further divided using the second-level clustering. Notably, the sub-clusters generated using different clustering methods exhibit good agreement based on the adjusted rand score and homogeneity score of 0.85 and higher.

B DETECTION OF HIGH VERSUS LOW WATER PRODUCERS USING SUPERVISED LEARNING

For purposes of supervised learning that follows the unsupervised learning, new features were extracted for each cluster based on the frequency of occurrence and depth-based distribution of the cluster along

the length of a well. Feature reduction was performed based on mutual information and p-value with thresholds of 0.2 and 0.08, respectively, for the Fort Worth Basin data and 0.05 and 0.15, respectively, for the Gulf Coast region data. The derived features were used to train KNN, logistic regression, and SVM supervised learning techniques to distinguish between the HWPs and LWPs. Additionally, due to availability of only a small size of well data from the basins, 100 iterations of K-fold cross-validation were performed to ensure the workflow is generalizable. This ensures that the generalization score of a supervised method developed on a small-sized dataset is not biased by the statistical differences due to the splitting of data into training and validation/testing dataset. The median Matthews correlation coefficient (MCC) and F-1 scores were used as performance metrics for the supervised methods. The performances of the regression models developed using various supervised learning methods are summarized in table 4. Each of the results are better than those published in our previous work that doesn't include the unsupervised learning step [22]. The pseudo-lithologies generated using the clustering enable better feature extraction that improves the identification of excess water-producing wells. Logistic regression performed the best for both the basins. Logistic regression performed better on the Fort Worth basin than the Gulf Coast basin. Performance of logistic regression is significantly better than KNN and support vector machine for the Gulf coast basin. The uncertainty in the performance quantified using the inter-quartile range (IQR) is much higher for the Gulf coast region compared to the Fort Worth region. Logistic regression has the overall lowest uncertainty in the performance.

TABLE 4: Performances of the new data-driven workflow when implementing various supervised-learning methods for the preemptive detection the high or low water producing wells. Interquartile range (IQR) quantifies the variation in the prediction performance.

Method	k-Nearest Neighbors				Logistic Regression				Support Vector Machine			
Score/ Metric	Median MCC	MCC IQR	Median F-1	F-1 IQR	Median MCC	MCC IQR	Median F-1	F-1 IQR	Median MCC	MCC IQR	Median F-1	F-1 IQR
Gulf Coast	0.64	0.10	0.67	0.10	0.93	0.16	0.95	0.11	0.71	0.20	0.71	0.14
Fort Worth	0.87	0.06	0.90	0.04	0.81	0.06	0.88	0.04	0.81	0.06	0.86	0.05

C DEPLOYMENT OF THE DATA-DRIVEN WORKFLOW TO PREEMPTIVELY DETECT THE HIGH-WATER PRODUCING WELLS

The trained and tested supervised models that exhibited high MCC and F1 scores were implemented in the data-driven workflow when deploying the workflow on new wells for the preemptive detection of whether the new well will be a high or low water producer. Such a deployment will first process a certain set of features extracted from the 5 well logs using KMeans and agglomerative clustering to generate the pseudo-lithology labels for each depth in the well. Following that, the deployment involves the processing of another set of features extracted from the 5 well logs based on the pseudo-lithologies assigned to each depth. The most generalizable regression model trained using supervised learning will be used to process these features to predict whether the new well will be a high or low water producer. The data-driven workflow uses 200 ft of data above the kick off point and 300 ft of data below the kick off point of a well. In doing so, the prediction of excess water production for a well serves as a preemptive detection for planning the reservoir/production management strategies.

IV Geophysical Signatures that Explain the Excess Water Production from the Unconventional Shale Wells

In this section, we discuss the newly discovered geophysical signatures that explain the excess water production from wells drilled in the unconventional shale reservoirs. Geophysical signatures were identified using Kendall Tau's test and permutation feature importance. Kendall Tau's test was used to quantify the strength of association between the continuous-valued features derived from the 5 logs based on the pseudo-lithology and the categorical target (HWP or LWP). Permutation feature importance was used to rank the features in accordance with the loss in the performance of the supervised learning models in the task of differentiating the HWP (high water producer) wells from LWP (low water producer) wells when the information of the feature is randomized. The strength of association of a feature with the target and the feature ranking were combined to discover the geophysical signatures that can explain the excess water production from the wells drilled in shales.

Recall that there are two major clusters, A and B, in the FW basin. Cluster A, generally, has much higher GR readings and is more permeable than cluster B. Notwithstanding, the LWPs are deeper, they are less permeable indicated by the smaller separation between the shallow and deep resistivity logs [23] and the NPHI and DPHI logs are generally – in all clusters – about 25 percent more separated in the LWPs than HWPs, signifying shalier and more clay-bearing rocks [24]; thus, low water production. Further, the gamma ray reading in all clusters for the LWPs are averagely 500 percent higher that for the HWPs. Especially, cluster B1 in LWPs have GR readings about 25 times than that in HWPs. The high GR readings in these organic-rich rocks indicates the presence of clays [25], which distinguish the LWPs from HWPs. Furthermore, statistical parameters derived from DPHI log from cluster A2 are 20% and 50% of the top ten feature ranking and association results, respectively. Cluster A, overall, is a porous, organic rich black shale common in FW basin. In the Barnett shale, DPHI log is considered a useful information to quantitatively assess shale gas resources [26]. More so, DPHI reading in LWPs is lower than that in HWPs. Furthermore, low DPHI values in shales implies low kerogen density [27,28], and low kerogen density is directly related to thermal maturity [29]. Jagadisan and Heidari suggested that kerogen at low thermal maturity could be water wet [30]. Therefore, the presence of cluster A2 with very low DPHI values can be considered one of the factors responsible for low water production in LWPs, since a water wet rock will tend to produce less water due to its strong affinity to water.

The presence of cluster D in a well was top ranked in both feature ranking and association for the GC basin. Cluster D seems to be a bituminous shale due to its average GR and high ILD readings; bituminous shales have excessively high TOC [31] and high resistivity in GC region suggests high TOC [32]. Further, cluster D occurs at a deeper depth in LWPs, thus the presence of bituminous shale in deep wells may signify an LWP. Additionally, the statistical parameters, mean and median, of the depth of cluster C occurred as top ranked and associated features. Cluster C appeared to be an averagely permeable pyrite-bearing shale considering its very low DPHI [33] and average GR reading. Cluster C is much deeper in LWPs than HWPs and can be considered as a signature distinguishing both classifications. The geological interpretations are consistent with what has been reported about the GC basin.

IV CONCLUSIONS

Data-driven workflow comprising unsupervised learning followed by supervised learning can be used to pre-emptively detect high and low water producing wells drilled in unconventional shale formations. Five conventional well log data from 29 wells in the Fort Worth basin and 23 wells in the Gulf Coast basin were used to develop and evaluate the data-driven workflow. The unsupervised learning for predicting the pseudo-lithology prior to well classification improves the pre-emptive detection of high and low water producing wells. The difference between neutron porosity and density porosity as well as the difference between logarithmic transformations of deep and shallow resistivity are important for assigning the pseudo-lithology. Average of neutron porosity and density porosity is an important feature for assigning the pseudo-lithologies for the wells from the Gulf-Coast region. Variation in the formulations of features used at various levels and clusters ensure reliable and robust multi-level clustering, which has a low bias from predominant pseudo-lithology.

The pseudo-lithologies generated using the clustering enable better feature extraction that improves the identification of excess water-producing wells. Logistic regression was the best supervised-learning technique for both the Gulf Coast and Fort Worth basins. For purposes of supervised learning, 393 features were generated for the Gulf Coast basin, while 337 features were generated for the Fort Worth basin. These features were reduced using univariate and bivariate statistical tests, such as Mutual Information and analysis of variance F-test. Such feature elimination or dimensionality reduction is essential to lower the variance and improve the generalization of the data-driven workflow. The uncertainty in the performance quantified using the inter-quartile range (IQR) is much higher for the Gulf coast region compared to the Fort Worth region. Logistic regression has the overall lowest uncertainty in the performance.

Kendall Tau's test and permutation feature importance method were used together to determine geophysical signatures that can explain the excess water production from the wells drilled in unconventional shales. Our analysis indicates that the low water producing wells intersect formations that exhibit higher clay content, shaliness, and lower permeability. Statistical parameters derived from DPHI log from cluster A2, which represent a porous black shale, are strong geophysical signatures that differentiate the high water-producing wells from the low water-producing wells. Low water-producing wells primarily contain the cluster A2 at deeper depths. Cluster A2 seems to be a shale with higher maturity and higher water wettability. In Gulf Coast basin, the presence of cluster D, a bituminous shale with high total organic carbon, is associated with differences between the low and the high water-producing wells. For the low water-producing wells, the cluster D occurs at deeper depths. Additionally, the mean and median of the depths of occurrence of cluster C, an averagely permeable pyrite-bearing shale, can be considered as a geophysical signature that distinguishes the high and the low water-producing wells.

Acknowledgement

We want to thank Berg-Hughes Center for Petroleum and Sedimentary Systems and Crisman Institute for Petroleum Research at Texas A&M University for providing financial support for the project. Also, we thank Mark Nibbelink and his team in Enverus, who have helped us by providing access to data and consultations on technical aspects of the Enverus (formerly Drillinginfo) platform.

V REFERENCES

[1] Kenneth Imo-Imo Eshiet, “Developments in the Exploitation of Unconventional Hydrocarbon Reservoirs,” in *Exploitation of Unconventional Oil and Gas Resources*, 2017, p. 13.

[2] U. S. E. I. A. (EIA), “World shale resource assessments,” 2015. [Online]. Available: https://www.eia.gov/analysis/studies/worldshalegas/.

[3] Y. Lei and J. Zhijun, “Global shale oil development and prospects,” *China Pet. Explor.*, vol. 24, no. 5, pp. 1–8, 2019.

[4] C. M. C Boyer, B Clark, V Jochen, R Lewis, “Shale gas: A global resource,” *Oilf. Rev.*, 2011.

[5] U.S. Geological Survey, “Assessment of Undiscovered Oil and Gas Resources of the,” 2004.

[6] D. M. Jarvie, B. Claxton, F. “Bo” Henk, and J. Breyer, “Oil and Shale Gas from Ft . Worth Basin, Texas,” *AAPG Natl. Conv.*, no. January 2001, pp. 1–28, 2001.

[7] R. J. Hill, D. M. Jarvie, J. Zumberge, M. Henry, and R. M. Pollastro, “Oil and gas geochemistry and petroleum systems of the Fort Worth Basin,” *Am. Assoc. Pet. Geol. Bull.*, vol. 91, no. 4, pp. 445–473, 2007, doi: 10.1306/11030606014.

[8] K. A. Bowker, “Barnett Shale gas production, Fort Worth Basin: Issues and discussion,” *Am. Assoc. Pet. Geol. Bull.*, vol. 91, no. 4, pp. 523–533, 2007, doi: 10.1306/06190606018.

[9] R. M. Pollastro, D. M. Jarvie, R. J. Hill, and C. W. Adams, “Geologic framework of the Mississippian Barnett Shale, Barnett-Paleozoic total petroleum system, Bend arch-Fort Worth Basin, Texas,” *Am. Assoc. Pet. Geol. Bull.*, vol. 91, no. 4, pp. 405–436, 2007, doi: 10.1306/10300606008.

[10] U.S. Energy Information Administration, “Drilling productivity report for key tight oil and shale gas regions,” 2020. https://www.eia.gov/petroleum/drilling/#tabs-summary-2 (accessed Aug. 12, 2021).

[11] Xinglai Gong., Y. Tian, D. A. McVay, W. B. Ayers., and J. Lee, “Assessment of the mexican eagle ford shale oil and gas resources,” *Soc. Pet. Eng. - SPE USA Unconv. Resour. Conf. 2014*, no. 2002, pp. 263–282, 2014, doi: 10.2118/168983-ms.

[12] B. R. Scanlon, R. C. Reedy, F. Male, and M. Walsh, “Water Issues Related to Transitioning from Conventional to Unconventional Oil Production in the Permian Basin,” *Environ. Sci. Technol.*, vol. 51, no. 18, pp. 10903–10912, 2017, doi: 10.1021/acs.est.7b02185.

[13] A. J. Kondash, E. Albright, and A. Vengosh, “Quantity of flowback and produced waters from unconventional oil and gas exploration,” *Sci. Total Environ.*, vol. 574, pp. 314–321, 2017, doi: 10.1016/j.scitotenv.2016.09.069.

[14] M. A. Engle *et al.*, “Origin and geochemistry of formation waters from the lower Eagle Ford Group, Gulf Coast Basin, south central Texas,” *Chem. Geol.*, vol. 550, no. June, p. 119754, 2020, doi: 10.1016/j.chemgeo.2020.119754.

[15] S. L. Montgomery, D. M. Jarvie, K. A. Bowker, and R. M. Pollastro, “Mississippian Barnett Shale, Fort Worth basin, north-central Texas: Gas-shale play with multi-trillion cubic foot potential,” *Am. Assoc. Pet. Geol. Bull.*, vol. 89, no. 2, pp. 155–175, 2005, doi: 10.1306/09170404042.

[16] M. M. Scales *et al.*, “A Decade of Induced Slip on the Causative Fault of the 2015 Mw 4.0 Venus

Earthquake, Northeast Johnson County, Texas," *J. Geophys. Res. Solid Earth*, vol. 122, no. 10, pp. 7879–7894, 2017, doi: 10.1002/2017JB014460.

[17] N. A. Khan, M. Engle, B. Dungan, F. O. Holguin, P. Xu, and K. C. Carroll, "Volatile-organic molecular characterization of shale-oil produced water from the Permian Basin," *Chemosphere*, vol. 148, pp. 126–136, 2016, doi: 10.1016/j.chemosphere.2015.12.116.

[18] Y. Jin and A. Davarpanah, "Using Photo-Fenton and Floatation Techniques for the Sustainable Management of Flow-Back Produced Water Reuse in Shale Reservoirs Exploration," *Water. Air. Soil Pollut.*, vol. 231, no. 8, 2020, doi: 10.1007/s11270-020-04812-7.

[19] A. Wilcox, "The water challenge program – Permian basin pilot results," *SPE/AAPG/SEG Unconv. Resour. Technol. Conf. 2018, URTC 2018*, 2018, doi: 10.15530/urtec-2018-2877246.

[20] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, "Application of machine learning and artificial intelligence in oil and gas industry," *Pet. Res.*, no. xxxx, 2021, doi: 10.1016/j.ptlrs.2021.05.009.

[21] J. H. Siddharth Misra, Hao Li, *Machine learning for subsurface characterization*. Gulf Professional Publishing, 2020.

[22] J. Foster, S. Misra, O. Osogba, and M. Bhatia, "Machine learning assisted detection of excess water-producing wells in unconventional shale plays," *J. Nat. Gas Sci. Eng.*, vol. 92, no. February, p. 104025, 2021, doi: 10.1016/j.jngse.2021.104025.

[23] S. J. Zarrouk and T. A. Moore, "PRELIMINARY ASSESSMENT OF THE GEOTHERMAL SIGNATURE AND ECBM POTENTIAL OF THE HUNTLY COALBED METHANE FIELD , NEW ZEALAND Geothermal Institute , Department of Engineering Science , University of Auckland , New Zealand Solid Energy New Zealand Ltd , P . O . ," no. December 2014, pp. 1–7, 2007.

[24] K. Bhuyan and Q. R. Passey, "Clay estimation from Gr and Neutron-density porosity logs," *SPWLA 35th Annu. Logging Symp. 1994*, 1994.

[25] T. Paronish, R. Toth, T. Carr, V. Agrawal, D. Crandall, and J. Moore, "Multi-Scale Lithofacies and Chemostratigraphic Analysis of Two Middle Devonian Marcellus Shale Wells in Northern West Virginia, USA," 2020, doi: 10.15530/urtec-2020-2763.

[26] Q. Fu *et al.*, "Log-derived thickness and porosity of the Barnett Shale, Fort Worth basin, Texas: Implications for assessment of gas shale resources," *Am. Assoc. Pet. Geol. Bull.*, vol. 99, no. 1, pp. 119–141, 2015, doi: 10.1306/07171413018.

[27] J. M. Salazar, R. J. M. Bonnie, W. W. Clopine, and G. E. Michael, "A practical petrophysical model for a source rock play: The Mancos Shale," *Interpretation*, vol. 5, no. 3, pp. T423–T435, 2017, doi: 10.1190/INT-2017-0014.1.

[28] R. Freedman, D. Rose, B. Sun, R. L. Brown, and T. Malizia, "Novel method for evaluating shale-gas and shale-tight-oil reservoirs using advanced well-log data," *SPE Reserv. Eval. Eng.*, vol. 22, no. 1, pp. 282–301, 2019, doi: 10.2118/181480-PA.

[29] K. S. Okiongbo, A. C. Aplin, and S. R. Larter, "Changes in type II Kerogen density as a function of maturity: Evidence from the Kimmeridge clay formation," *Energy and Fuels*, vol. 19, no. 6, pp. 2495–2499, 2005, doi: 10.1021/ef050194+.

[30] A. Jagadisan and Z. Heidari, “Demystifying wettability alteration in kerogen as a function of its geochemistry and reservoir temperature and pressure using molecular dynamics simulations,” *Proc. - SPE Annu. Tech. Conf. Exhib.*, vol. 2019-Septe, 2019, doi: 10.2118/195863-ms.

[31] A. K. Singh *et al.*, “Geochemical and organic petrographic characteristics of high bituminous shales from Gurha mine in Rajasthan, NW India,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–19, 2020, doi: 10.1038/s41598-020-78906-x.

[32] O. Ogiesoba and U. Hammes, “Seismic-attribute identification of brittle and TOC-rich zones within the Eagle Ford Shale, Dimmit County, South Texas,” *J. Pet. Explor. Prod. Technol.*, vol. 4, no. 2, pp. 133–151, 2014, doi: 10.1007/s13202-014-0106-1.

[33] C. C. A. H. C. Scala, “Effect of Pyrite on Resistivity and Other Logging Measurements,” in *Paper presented at the SPWLA 17th Annual Logging Symposium, Denver, Colorado*, pp. 1–34.

VI APPENDIX A: ADDITIONAL FIGURES

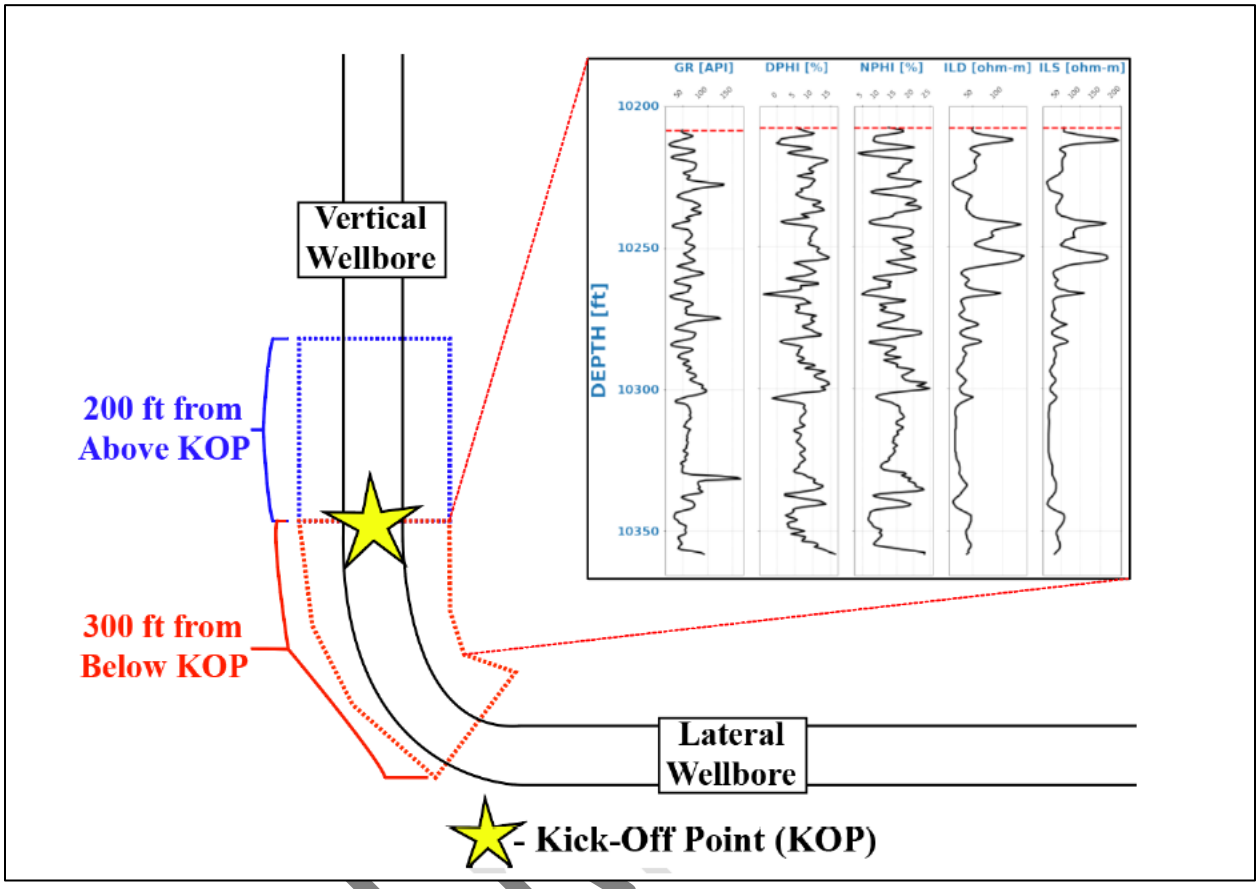

Figure A1: Five well logs from 200 feet above the kick-off point and 300 feet below the kick-off point were used for the desired pre-emptive detection of wells producing excess water relative to oil.

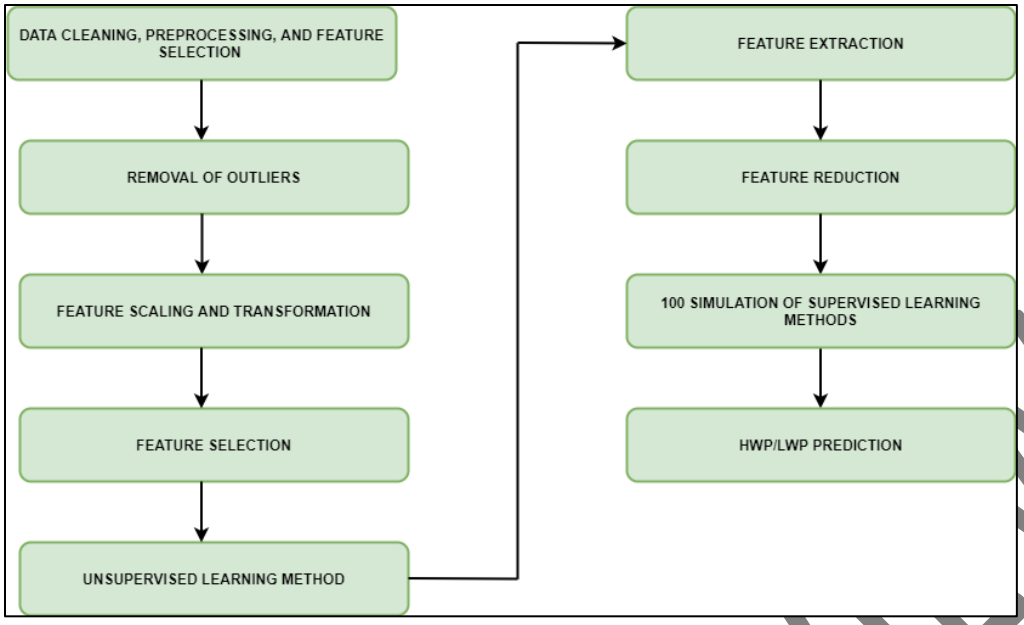

Figure A2: The data-driven workflow, based on unsupervised learning followed by supervised learning, for the pre-emptive detection of high versus low water producing wells drilled in unconventional shale formations.

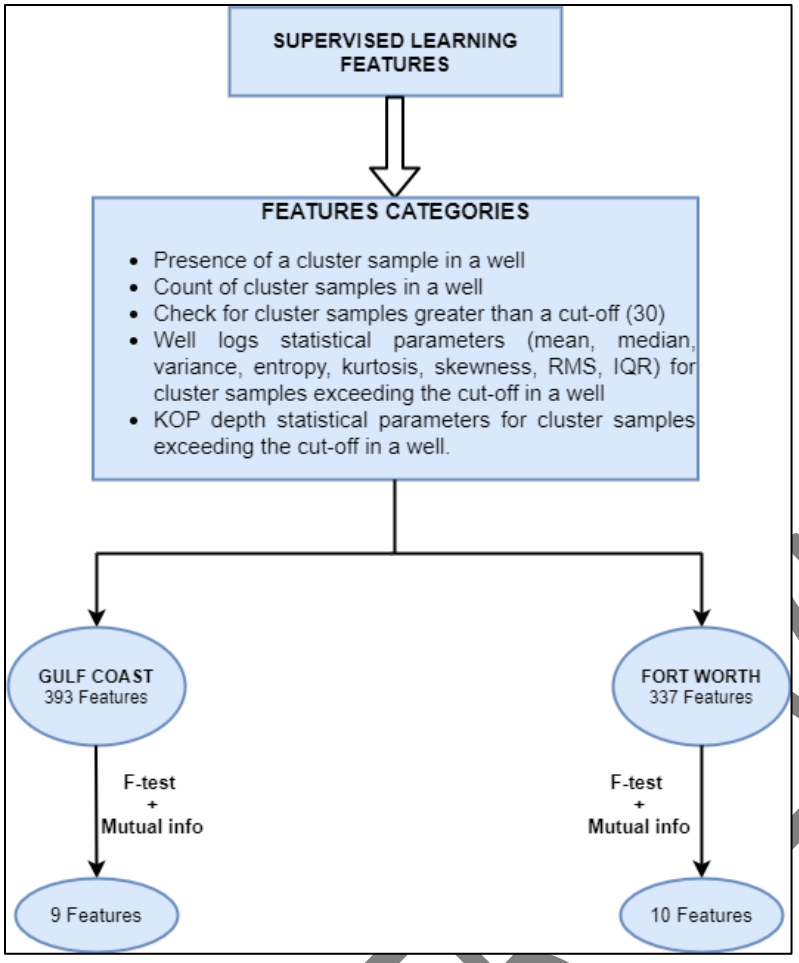

Figure A3: Feature extraction followed by feature reduction required for the development of highly generalizable supervised learning.