InterPore JOURNAL

Original Research

MACHINE LEARNING-BASED ESTIMATION OF OIL RECOVERY FACTOR USING XGBOOST: INSIGHTS FROM CLASSIFICATION AND DATA-DRIVEN ANALYSES

Alireza Roustazadeh[1], Frank Male[2], Behzad Ghanbarian[3,4,5*], Mohammad B. Shadmand[6], Vahid Taslimitehrani[7], Larry W. Lake[8]

[1]Porous Media Research Lab, Department of Geology, Kansas State University, Manhattan, Kansas, United States; [2]John and Willie Leone Family Department of Energy and Mineral Engineering, Pennsylvania State University, University Park, Pennsylvania, United States; [3]Department of Earth and Environmental Sciences, University of Texas at Arlington, Arlington, Texas, United States; Department of Civil Engineering, University of Texas at Arlington, Arlington, Texas, United States; [5]Division of Data Science, College of Science, University of Texas at Arlington, Arlington, Texas, United States; [6]Department of Electrical and Computer Engineering, College of Engineering, University of Illinois at Chicago, Chicago, Illinois, United States; [7]Machine Learning/Data Science Tech Lead at Meta, San Francisco, California, United States; [8]Hildebrand Department of Petroleum and Geosystems Engineering, University of Texas at Austin, Austin, Texas, United States

Correspondence to:
Behzad Ghanbarian at ghanbarianb@uta.edu

How to Cite:
Roustazadeh, A., Male, F., Ghanbarian, B., Shadmand, M. B., Taslimitehrani, V., & Lake, L. W. (2025). Machine Learning-Based Estimation of Oil Recovery Factor Using XGBoost: Insights from Classification and Data-Driven Analyses. *InterPore Journal*, 2(3), IPJ250825-4. https://doi.org/10.69631/ipj.v2i3nr53

RECEIVED: 24 Nov. 2024
ACCEPTED: 19 May 2025
PUBLISHED: 25 Aug. 2025

ABSTRACT

In petroleum engineering, it is essential to determine the ultimate recovery factor (RF) particularly before exploitation and exploration. However, accurately estimating requires data that may not be necessarily available or measured at early stages of reservoir development. To rectify this, we applied machine learning (ML) to estimate oil RF from readily available features. To construct the ML models, we applied the XGBoost classification algorithm. Classification was chosen over regression because recovery factor is bounded from 0 to 1, much like probability. Three databases with various reservoir properties and recovery factors were used, leaving us with four different combinations to first train and test the ML models and then further evaluate them using an independent database including unseen data. Cross-validation with ten folds was applied on the training datasets to assess the effectiveness of the models. To evaluate the accuracy and reliability of the models, the accuracy, within-1 accuracy, precision, recall, macro-averaged f1 score and R^2 were determined. Overall, results showed that the XGBoost classification algorithm could estimate the RF class with accuracies as high as 0.77 in the training datasets, 0.36 in the testing datasets and 0.24 in the independent databases used. We found that the reliability of the XGBoost classification model depended on the data in the training dataset, indicating that the ML models were database dependent. The feature importance analysis and the Shapley Additive exPlanations (SHAP) approach showed that the most important features were reserves, reservoir area and thickness.

KEYWORDS
Classification, Machine learning, Oil recovery factor, Extreme gradient boost

1. INTRODUCTION

Accurate estimation of the ultimate recovery factor (RF) has broad applications to oil and gas exploration and carbon storage. The RF, when combined with the reservoir size and production costs, indicates whether a reservoir would be economical if developed (26). Accordingly, various methods—such as dynamic reservoir simulations, production decline curve analysis, material balance, and field analogues (4, 50, 55)—have been proposed over the past several decades to determine the RF from available measured data. Such methods, however, are either computationally demanding, associated with significant uncertainties and errors, or require input data that may not be readily available prior to field development (41, 43, 57, 72). With recent advances in artificial intelligence, machine learning (ML) algorithms have been used over the past several years to replace these methods. The goal of this study is to train ML models to estimate the RF using readily available and easily measurable data. Such models could offer a robust and cost-effective tool to support decision-making in identifying viable projects for development (29, 53, 66).

Machine learning models have recently gained significant popularity and have been successfully applied across various domains of petroleum engineering (54, 20). In the literature, most ML models developed to estimate RF are based on regression analysis (2, 4 , 5, 6, 39, 64, 69). For example, Srivastava et al. (66) applied ML to the 2013 version of the Atlas of Gulf of Mexico database to cluster similar reservoirs and subsequently performed regression to estimate the RF. First, they conducted a principal component analysis to reduce the dimensionality of the original features in their database. Next, they used k-means clustering to group similar reservoirs, and finally, they employed partial least squares regression to correlate input features with the RF. Those authors found that the combination of k-means clustering, principal component analysis, and partial least square regression led to unsatisfactory results, achieving a maximum Pearson's correlation coefficient of 0.2 between predictions and realized RF. However, when they instead grouped reservoirs using dimensionless numbers—such as gravity number, aspect ratio, and density number—and used these numbers as input features for their partial least squares regression model, they achieved better results. The partial least squares regression and k-means clustering based on dimensionless numbers method yielded R^2 values that ranged from 0.92 (excellent) to 0.1 (very poor) in four different clusters.

Kaczmarczyk et al. (30) applied clustering analysis and tree-based regression on data from three different databases (TORIS, Digital Knowledge System, and the Oil and Gas Journal). They estimated primary, secondary, and tertiary RFs. Their decision tree model was constructed using six different reservoir and fluid properties (i.e., pressure, permeability, viscosity, porosity, API gravity, and depth). Decision trees such as these are very prone to overfitting and have weak generalizability.

Gupta et al. (24) applied partial least squares regression to data from deep offshore assets in the Gulf of Mexico to estimate the RF variance between the early appraisal phase and the post sanction phase. Their work does not directly estimate RF. Instead, it forecasts the difference in predicted RF from two different stages in the development process. In another study, Karacan (33) estimated tertiary RF using fuzzy logic, reporting an R^2 of 0.88. Unfortunately, he used reservoir data collected from only 24 reservoirs.

Recently, Roustazadeh et al. (61) developed regression-based models using three ML algorithms: XGBoost, support vector machines, and forward stepwise multiple regression. They found that the XGBoost regression model was slightly more accurate than the other two models. Roustazadeh et al. (61) also reported that the results of XGBoost regression model were database-dependent. Pooladi-Darvish et al. (60) also demonstrated the accuracy of the XGBoost model in estimating oil recovery factor

using data from 18,000 reservoirs, based on input features such as pressure, porosity, permeability, API gravity, temperature, viscosity, depth, thickness and water saturation. Using synthetically generated data, Matkerim et al. (51) showed that the accuracy of the XGBoost model in estimating oil RF was comparable to that of random forest and neural network models.

To the best of our knowledge, classification-based ML models have not yet been used to estimate oil RF at the reservoir scale using a large database comprising several thousand samples. As shown above, regression algorithms have been previously applied in the literature (13, 25, 31, 48). Furthermore, the majority of ML models have not undergone further evaluation using new, unseen data that was not part of their training. Although Roustazadeh et al. (61) showed that the performance and predictions of the XGBoost regression model were dependent on underlying databases, similar analyses for classification models in the context of oil recovery estimation remain unaddressed. Therefore, the main objectives of this study are to: 1) apply extreme gradient boost (XGBoost) classification to construct ML models using large databases, 2) estimate the ultimate oil RF at the early stages, and 3) investigate the database dependence of classification-based ML models by evaluating them using unseen data.

2. MATERIALS AND METHODS

In this section, we first explain the data used in this study. Next, we describe the data preparation process for constructing the ML-based models using the XGBoost classification algorithm. Finally, we explain how Shapley Additive exPlanations (SHAP) was used to the determine feature importance.

2.1. Databases

The data used in this study come from the following three databases, which are briefly described below.

2.1.1. Commercial database

The private, commercial database contains information on over 1200 conventional reservoirs worldwide, encompassing a variety of rock and fluid properties. It includes more than 200 features per reservoir, such as pressure, water saturation, and porosity. This database provides data for both oil and gas reservoirs. After applying filtering criteria, we selected data from 600 oil fields for use in this study. For additional details on this database, please refer to Lee and Lake (39).

2.1.2. Tertiary Oil Recovery Information System database

The Tertiary Oil Recovery Information System (TORIS), compiled by the National Petroleum Council and used by the US Department of Energy, is a well-established and respected database containing information on oil reservoirs across the United States of America. It includes approximately 1300 observations and 60 features; however the database also has a relatively high number of missing values. For further details, see Long (45).

2.1.3. Bureau of Ocean Energy Management Atlas of the Gulf of Mexico

The Bureau of Ocean Energy Management (BOEM) annually collects oil and gas data from conventional fields on the outer shelf of the Gulf of Mexico (8). In this study, we used the 2018 version of this database, hereafter referred to as Atlas. The Atlas database contains over 13,000 observations and 80 features covering both oil and gas reservoirs. For the purpose of this study, only data from oil fields were used to construct the ML models. Among the three databases analyzed, Atlas has the lowest proportion of missing values.

Reservoirs appearing in multiple databases were de-duplicated by removing duplicate entries. The observation with fewer input features was removed. **Figure 1** shows the distributions of the RF, porosity (ft^3/ft^3), and natural logarithm of permeability in the Commercial, Atlas, and TORIS databases. The distributions of RF, permeability, and porosity in the TORIS and Commercial databases are similar, while the Atlas database is different. The distribution of natural logarithm of permeability is left-skewed and heavy-tailed for the TORIS and Commercial databases, while slightly left-tailed for the Atlas database. The RF distributions for all the three databases are right-skewed with heavier tails in the TORIS and Commercial databases.

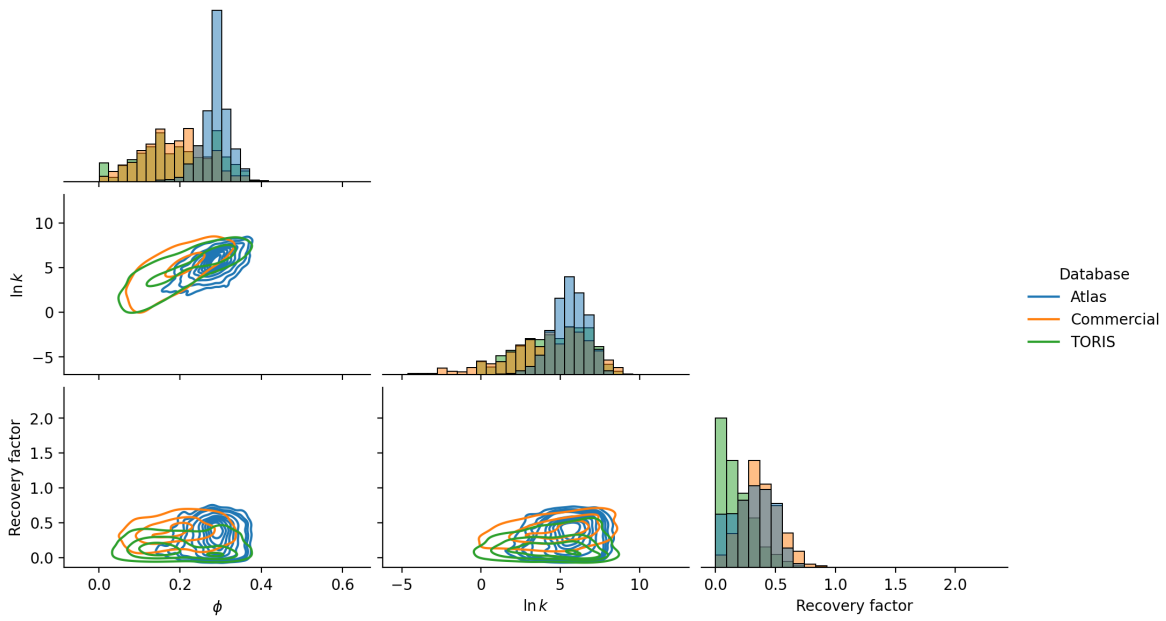

Figure 1: Distributions of the oil recovery factor (dimensionless), porosity, and natural logarithm of permeability (in mD) for databases used in this study.

Figure 2 shows letter-value (extended box and whisker) plots for each feature of interest in each database used in this study prior to data processing. As can be seen, the databases contain a few outliers in porosity, permeability, and RF values. The TORIS and Commercial databases display similar interquartile ranges, as well as comparable median, and mean values. In contrast, the Atlas database consists of samples with smaller ranges in porosity and permeability compared to those in the TORIS and Commercial databases. It also exhibits considerably different mean and median values compared to TORIS and the Commercial database, as shown in all three plots. TORIS had the greatest number of RF outlier values, while Atlas had the fewest among the three databases. The oil RF values have statistically different averages between datasets.

The Commercial, TORIS, and Atlas databases described above were used to construct classification models for estimating oil RF based on other reservoir properties. These databases were merged using four different combinations to create larger databases. More specifically, the combined TORIS and Commercial databases were labelled **TC**, TORIS and Atlas were labelled **TA**, and Commercial and Atlas were labeled **CA**. The last combination was created by merging all the three databases (i.e., TORIS, Commercial, and Atlas) and labelling it **TCA**. In the first three combinations (i.e., TC, TA, and CA), the third database was used for testing.

2.2. Data preparation and feature engineering

Data preparation began by removing any reservoirs from the databases that did not have a recorded RF value. Although many different geological, petrophysical, and production features were available, we selected only those features that were available in all three databases. Moreover, any feature associated with post exploration phase data (e.g., production time, final abandonment pressure, and cumulative production) was removed. **Appendix A** (available for download online here) presents heat map matrices based on Spearman's rank correlation coefficients calculated for eleven input features and one output feature across the different databases.

To conserve data similarity across all databases, the selected features were constrained to the following ranges: the oil formation volume factor $1 \leq B_0 \lesssim 3$ (34), gas oil ratio $0 \leq GOR \leq 60$ Mscf/stb (61), and reserves between 0 and 5×10^{11} stock tank barrels (14)—the Atlas database defines reserves as the hydrocarbon remaining that could be economically recovered. We should point out that by reserves we mean original oil in place. Any feature with more than 70% missing values and any reservoir with more

than 55% missing values were removed. **Table 1** lists the final input features and target variable (RF) as well as their ranges in each database.

For the purpose of constructing ML classification models, ten RF classes were defined using intervals of 0.1. For example, samples with RF values between 0 and 0.1 were assigned to class 0. Similarly, samples with 0.9≤ RF ≤1 were in class 9. We randomly selected 90% of the data for training and cross-validation and 10% for testing, similar to the study by Dias et al. (18). The purpose of training and testing splits is to provide the ML models with enough training data to be constructed upon and avoid over-fitting (1). We used 10-fold cross-validation to tune the model hyperparameters on the training dataset (9). We split the data into the training and testing sets prior to any data preprocessing or imputation. Then, we

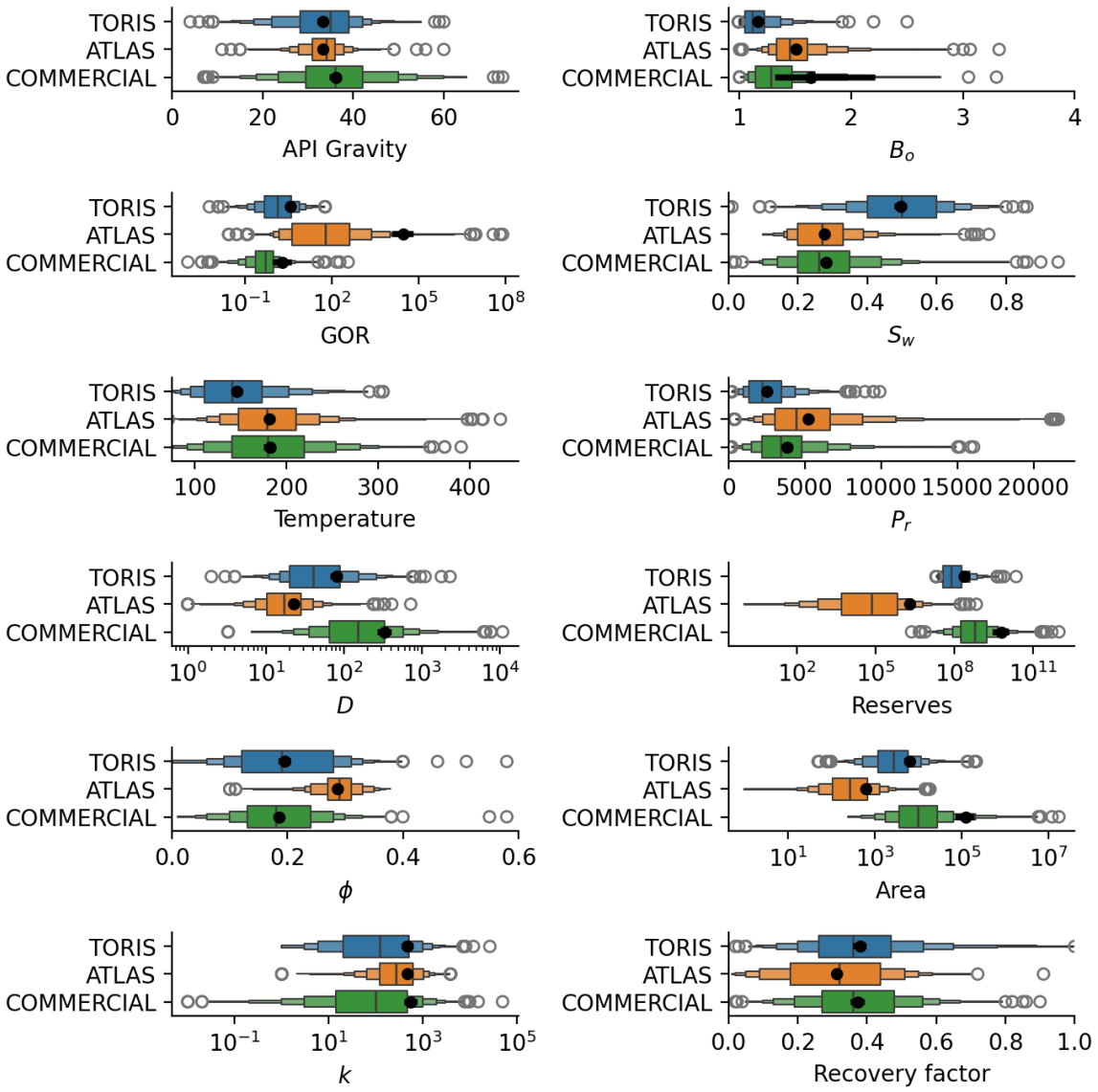

Figure 2: Letter-value plots of the 11 input features and 1 output feature for the three databases used in this study before data processing and preparation. ***Units are as follows:*** *degrees API for gravity; bbl/rb (barrels/reservoir barrel) for Bo (Bo = oil formation volume factor); Mscf/stb (thousand standard cubic feet per stock tank barrel) for GOR (gas oil ratio); dimensionless for water saturation (S_w); F° for temperature; psi for P_r (pressure); D = thickness; bbl (=?) for reserves; dimensionless for porosity (ø); acres for area; millidarcy for k; and dimensionless for recovery factor.* Dots in each box represent the mean of the distribution, with bootstrapped uncertainty in the mean. The lines inside the boxes show the median of the distribution. The lower boundaries in each box show the 25th quantile and the upper boundaries of the boxes show the 75th quantile. Extended boxes show the 12.5th percentile, 6.25th percentile, etc. Circles represent the outliers in each distribution.

 https://doi.org/10.69631/ipj.v2i3nr53

Table 1: Final input features and target variable (RF) as well as their ranges within each generated database. Recall that TC refers to the combination of TORIS and Commercial, TA to TORIS and Atlas, CA to Commercial and Atlas, and TCA to the merged dataset containing TORIS, Commercial, and Atlas.

Feature	Database TC	Database TA	Database CA	Database TCA
API Gravity	4-62	4-60	7-62	4-62
B_o (RB/STB)	0.99-3.05	0.99-3.32	1-3.32	0.99-3.32
GOR* (MSCF/RB)	0.001-57	0.006-59.62	0.001-59.62	0.001-59.62
Water saturation, S_w (-)	0-0.86	0-0.86	0.01-0.86	0-0.86
Temperature (°F)	43-360	47-305	42.8-360	42.8-360
Pressure, P_r (psi)	140-12,820	140-21609	200-21609	140-21609
Thickness, D (ft)	2-7681	1-2300	1-7681	1-7681
Reserves (STB)	2.33×10^{6}-4.99×10^{11}	1-2.2×10^{10}	1-4.99×10^{11}	1-4.99×10^{11}
Permeability (mD)	0-5×10^{4}	0-2.6×10^{4}	0.01-5×10^{4}	0-5×10^{4}
Porosity, ϕ (ft^3/ft^3)	0-0.58	0-0.58	0.01-0.58	0-0.58
Area (acre)	50-6.5×10^{6}	1-1.4×10^{5}	1-6.5×10^{6}	1-6.5×10^{6}
Oil RF	0.02-1	0.01-1	0.01-0.91	0.01-1

API = American Petroleum Institute; B_o = oil formation volume factor (function of temperature and pressure); GOR = Gas Oil Ratio; MSCF/RB = Thousands Standard Cubic feet per Reservoir Barrel; STB = Stock Tank Barrel; RF =

preprocessed the testing data with the identical parameters used on the training dataset (32). We did not perform any imputation on the independent databases. Instead, we removed samples with missing data among either input features or the target variable. This is because the independent databases were used to further evaluate the ML models using real-time data, not imputed ones. As stated earlier, observations in the training and testing datasets were de-duplicated. To ensure the fraction of samples in each class for the training and testing datasets was approximately equal, both the training and testing datasets were stratified by recovery factor class (7, 28).

2.3. Machine learning and XGBoost classification model

The aim of this study was to apply an ML classification model to estimate an expected RF interval for a reservoir at the exploration stage. We used the XGBoost model, an open-source algorithm with a Python API. XGBoost may be used for either classification (27, 59) or regression (19, 56) problems. It is an

Table 2: Hyperparameters and their optimized values for the XGBoost model developed on four combinations of three databases. Recall that TC is TORIS merged with commercial, TA is TORIS merged with Atlas, CA is Commercial merged with Atlas, and TCA is the combination of TORIS, commercial and Atlas. The training datasets in the TC, TA, CA, and TCA databases consisted of 1669, 5311, 5172, and 6076 samples, respectively.

Parameters	Database TC	Database TA	Database CA	Database TCA
Max depth	2	4	7	7
Minimum child weight	6	11	8	9
Learning rate	0.21	0.30	0.30	0.30
Subsample	0.94	0.70	0.74	0.98
Column sample by tree	0.59	0.91	0.99	0.90
Objective	Multi:softmax	Multi:softmax	Multi:softmax	Multi:softmax
Evaluation metric	mlogloss	mlogloss	mlogloss	mlogloss
Alpha	0.2	0.2	0.2	0.2
Lambda	0.01	0.01	0.01	0.01
Column sample by level	0.9	0.9	0.9	0.9
Gamma	0.01	0.01	0.01	0.01
Max delta step	0.1	0.1	0.1	0.1
Number of classes	10	10	10	10

ensemble ML algorithm rooted in decision trees, but it mitigates the risk of overfitting inherent in tree-based methods by using gradient boosted on model errors (12). In the XGBoost framework, weak learners are stacked to create strong learners (68).

XGBoost can provide more accurate estimations than other ML models, such as support vector machines (44, 59), random forest and k-nearest neighbors (11, 27). In addition, gradient boosting algorithms (e.g., XGBoost) are unaffected by collinearity among input features (10, 36, 52).

The tunning of hyperparameters is required to develop more accurate ML-based models (37). During hyperparameter tuning, we optimized hyperparameters with 10-fold cross-validation, using ray-tune with hyperopt to select the models with the lowest multiclass logistic (logarithmic) loss function (mlogloss). The hyperparameter values that yielded the lowest value of mlogloss were used to construct the ML models.

2.4. Model evaluation

The performance and accuracy of the models were evaluated using four parameters including accuracy, within 1 class accuracy, macro averaged f1 score and R^2, as given by **Equations 1** to **6**,

$$accuracy = \frac{Correct\ classifications}{Total\ classifications} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$within\ 1\ accuracy = \frac{Number\ of\ estimations\ \in Neighbour\ classes}{Total\ number\ of\ estimations} \quad (2)$$

$$precision = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4)$$

$$f1\ score = \frac{\sum_{i=1}^{n} class\ i's\ f1\ score}{n} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{j=1}^{N}\left(RF_j^{meas} - RF_j^{est}\right)^2}{\sum_{j=1}^{N}\left(RF_j^{meas} - \langle RF_j^{meas}\rangle\right)^2} \quad (6)$$

where n is the number of classes, N is the number of samples, $precision$ is the proportion of positive classifications that are actually positive, $recall$ is the proportion of actual positives that are correctly identified and TP, TN, FP and FN represent the true positives, true negatives, false positives and false negatives, respectively. All these parameters range from 0 and 1. Within 1 class accuracy indicates the number of estimations in one class above and/or below the desired class.

2.5. Shapley Additive exPlanations and Permutation Feature Importance

Shapley Additive exPlanations, commonly known as SHAP, can estimate the importance of each feature and its impact on ML model outputs. It was originally proposed by Shapley (63) who applied concepts from game theory to determine the outcome of a game based on every individual player's input. It explains how each feature influences a ML model (46). This method (SHAP) has been widely applied to determine feature importance (35, 40, 49).

Permutation feature importance is another useful measure of feature importance. In this method, a feature is randomly permuted several times, and the model is used to predict the target variable. Its permutation feature importance is how much, on average, the model has been degraded by garbling that input. For a feature with high permutation feature importance, the model is much worse after permuting that feature. For one with a low permutation feature importance, the model is unaffected or even (though this is likely) improved.

Both methods, SHAP and Permutation Feature Importance, offer slightly different roles. Shapley Additive exPlanations explain how a model arrives at a prediction. It is useful when examining the data a model

was trained on. Permutation Feature Importance, on the other hand, is more useful when looking at data that the model was not trained on. We use SHAP on the training data for each model and Permutation Feature Importance on the testing data and independent dataset.

2.6. Workflow of the constructed models

Figure 3 presents the workflow used for constructing the ML models, starting with data selection and de-duplication on the databases used in this study including TC (TORIS merges with Commercial), TA (TORIS merged with Atlas), CA (Commercial merged with Atlas), and TCA (the combination of all three databases). Next, every model was constructed using XGBoost to estimate the oil RF classes in the training and testing datasets. They were then evaluated on independent databases. In total, four ML models were constructed and evaluated using accuracy, neighborhood accuracy, and the macro average f1 score.

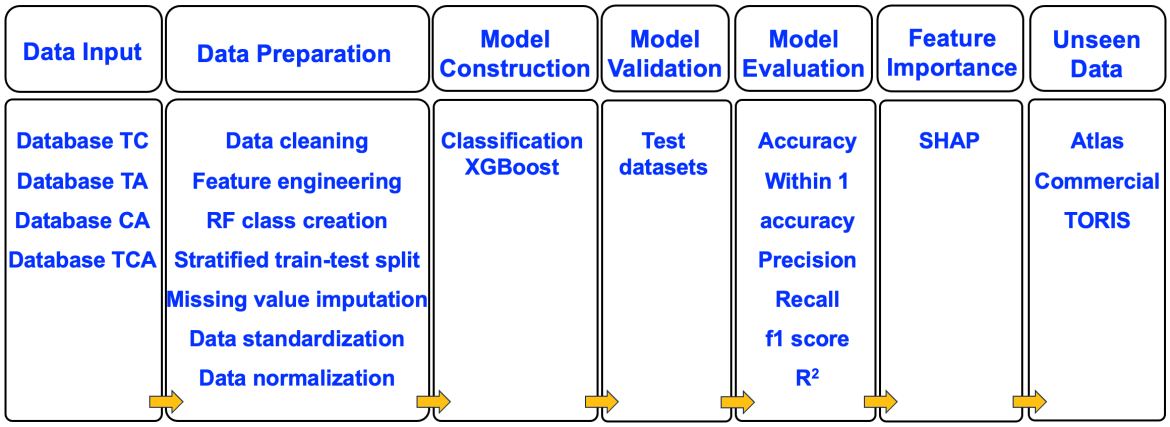

Figure 3: The workflow used to prepare the data, construct the models and evaluate the models using independent databases.

3. RESULTS

3.1. Hyperparameter tuning

Table 2 lists the optimized hyperparameters used to construct the classification-based ML models using the XGBoost algorithm. As the number of training samples increased—from 1669 samples in the TC database to 6076 samples in the TCA database—the hyperparameters became easier to optimize. The optimum learning rate decreased as the training dataset increased. This means that the best performing models took more steps to reach the optimum value that had the minimum overfitting or underfitting.

3.2. Oil RF estimation

Table 3 presents the calculated values for accuracy, within-1 accuracy, precision, recall, macro-averaged f1 score, and R^2 for both the training and testing datasets, as well as for the independent databases used to further evaluate the constructed ML models. The number of samples in the training datasets were 1669 for TC, 5311 for TA, 5172 for CA, and 6076 for TCA. The lowest accuracy and f1 score values (i.e., 0.31 and 0.29) belong to the database TC, which had the fewest observations. An accuracy of 0.31 means that the XGBoost model predicted the correct oil RF class for 31% of the samples. The f1 score of 0.29 is the harmonic average of the precision and recall for the model trained on TC. Ideally, all metrics would be 1 (23, 40).

As the number of samples in the training dataset increases from the database TC to TA, the accuracy and f1 score values also increase (accuracy = 0.31 vs. 0.34 and f1 score = 0.29 vs. 0.32). Similar results were observed for the CA and TCA databases, which indicates the ML models improve as they are given more data. Although the TCA database contains nearly 700 more samples than the TA database, its accuracy and f1 score values (0.36 and 0.35) are not greatly improved from those reported for TA (0.34 and 0.32). However, the RF estimations for the TCA and TA databases are more accurate than those for

Table 3: The performance metrics for the XGBoost classification algorithm employed to predict the oil RF class in each database used in this study.

Database	Metric	Train	Test	Independent
TC	accuracy	0.48	0.31	0.20
	within-1 accuracy	0.75	0.75	0.55
	precision	0.51	0.33	0.14
	recall	0.48	0.31	0.20
	f1 score	0.45	0.29	0.13
	R^2	0.23	0.16	-0.49
TA	accuracy	0.59	0.34	0.24
	within-1 accuracy	0.81	0.73	0.68
	precision	0.59	0.33	0.20
	recall	0.59	0.34	0.24
	f1 score	0.58	0.32	0.20
	R^2	0.44	0.22	-0.02
CA	accuracy	0.75	0.35	0.24
	within-1 accuracy	0.87	0.75	0.62
	precision	0.76	0.33	0.20
	recall	0.75	0.35	0.24
	f1 score	0.75	0.33	0.20
	R^2	0.59	0.28	-0.31
TCA	accuracy	0.77	0.36	-
	within-1 accuracy	0.88	0.75	-
	precision	0.77	0.36	-
	recall	0.77	0.36	-
	f1 score	0.76	0.35	-
	R^2	0.62	0.32	-

the CA database in the training. We found the testing f1 score values were consistently lower than the training f1 values for all databases.

The within-1 class accuracy is also reported in **Table 3**. For the training dataset, we found the within-1 accuracy varied from 0.75 to 0.88. For the testing dataset, this dropped by about 10 percentage points for all databases except for TC, which remained constant. These values show that the constructed models estimate the oil RF with approximately 75% accuracy within the correct or a neighboring class with in-sample data. Given that the interval in the RF classes is 0.1 (10%), the obtained results indicate reasonable estimations by the XGBoost model.

The ML models estimated the RF less accurately for the independent databases (unseen data) than for the testing datasets, aligning with Roustazadeh et al. (61), who highlighted the database dependence of regression-based ML models. This means that ML models are more accurate if the testing and training data are statistically similar. As **Table 3** shows, accuracy fell into the 0.2 range, and R^2 went to or below zero for these out-of-sample databases. Near zero and negative R^2 values were also reported by Kumar et al. (37) who developed regression-based ML models to estimate RF for reservoirs within the TORIS and Gulf of Mexico databases (see Table 3 and Table 5 in 37). The R^2 values reported in **Table 3** are also smaller than those reported by Thanh et al. (71) who applied general regression neural network, cascade forward neural network with Levenberg-Marquardt optimization, cascade forward neural network with Bayesian regularization, and XGBoost models to estimate RF. They reported an R^2 greater than 0.97 for both training and testing datasets (see Table 5 in 71). This is because of lack of diversity in their database,

which contained only 260 samples. More specifically, their input features—permeability, initial oil in place saturation, total pore volume and injected pore volume—approximately followed a uniform distribution (see Fig. 12 in 71).

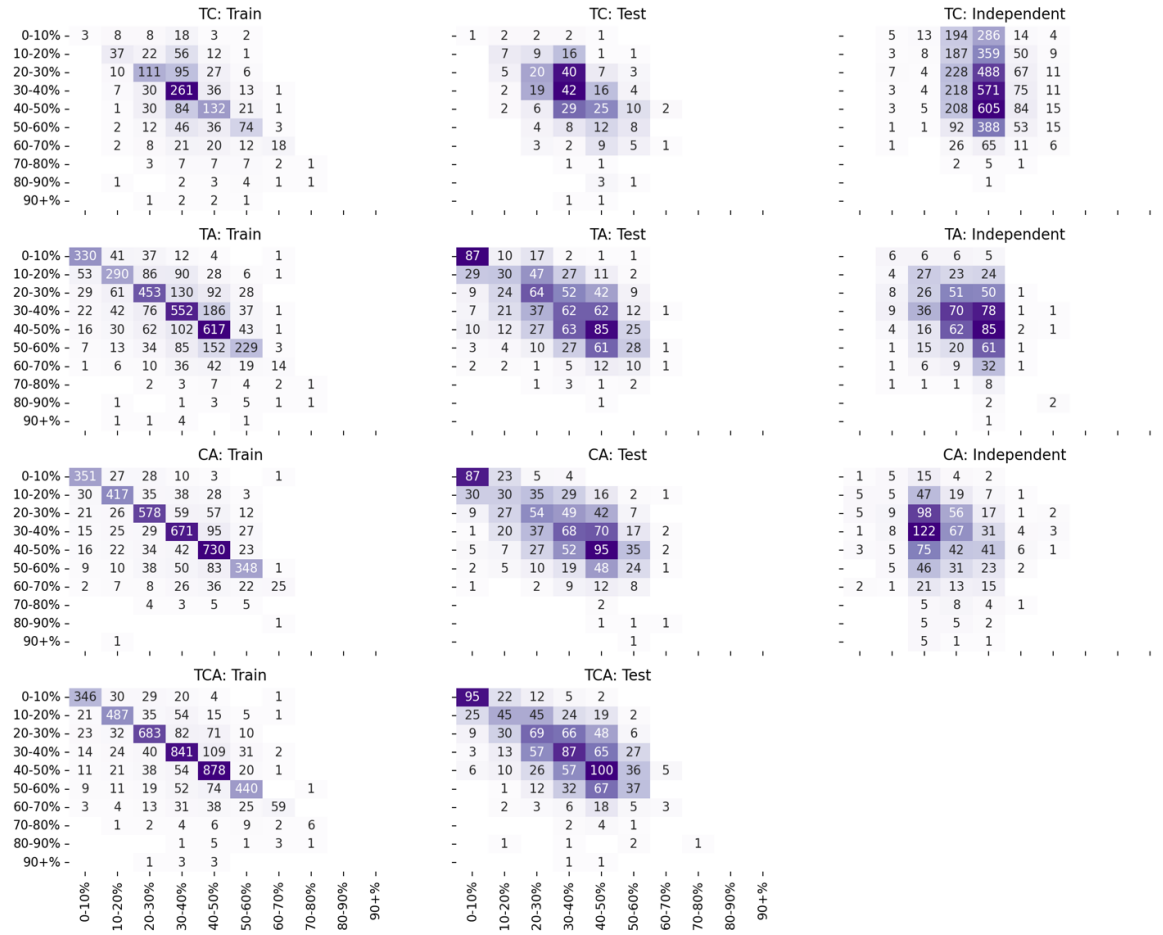

Figure 4: Confusion matrices for true RF classes (left) compared to XGBoost predictions (bottom axis) for the training and testing datasets in addition to the independent database. The number in each cell indicates the number of samples in that RF class. TC denotes TORIS merged with Commercial, TA is TORIS merged with Atlas, CA represents Commercial merged with Atlas, and TCA denotes Commercial merged with TORIS and Atlas.

Heat maps showing the estimated versus actual oil RF in each class for the four database combinations are shown in **Figure 4**. For the training data, the model performs well, following the diagonal. However, for the testing plots, the constructed models tend to overestimate the oil RF for the lower classes and to underestimate it for the higher classes, in accord with the results of Lee and Lake (39), Talluru and Wu (67), Makhotin et al. (48), and Roustazadeh et al. (61). **Figure 4** shows that most samples have an RF value between 0.2 and 0.5, corresponding to intermediate classes (i.e., 4, 5, and 6).

Figure 4 also shows that most training and testing observations are either on or adjacent to the long diagonal. This is consistent with the high within-1 accuracies (>0.74) reported in **Table 3**. Particularly for 0.2< RF <0.5, the performance of the constructed ML models is satisfactory. This range corresponds to that of most conventional oil reservoirs (22, 38, 60), indicating the potential practical applicability of the model to such reservoir types. Conventional reservoirs have an average RF of approximately 0.35 (65). While the models generally succeed in predicting low-RF for low-RF reservoirs, their performance becomes less reliable when predicting high-RF reservoirs.

 https://doi.org/10.69631/ipj.v2i3nr53

3.3. Feature importance

Feature importance analysis was performed for the four combinations of databases studied here. **Figure 5** shows the results of the SHAP feature importance analysis on the training data and permutation feature importance on the testing and independent datasets. As can be observed, reserves, reservoir thickness, and area are consistently the top features in the ML models, consistent with the results of Roustazadeh et al. (61) who estimated oil and gas RF using the regression-based ML models. Their feature importance analyses in general are similarly ranked, but the drop-off between reserves and other features varies significantly. For the features reservoir thickness and area, one may argue that larger reservoirs are more probable to have higher oil RFs. Mahmoud et al. (47) collected data from 173 reservoirs and applied an artificial neural network to train and test a predictive model using 77% and 23% of their data. They identified reservoir area as one of the most important features and observed a positive correlation between oil RF and reservoir area (see their Fig. 1 and Table 2 in 47).

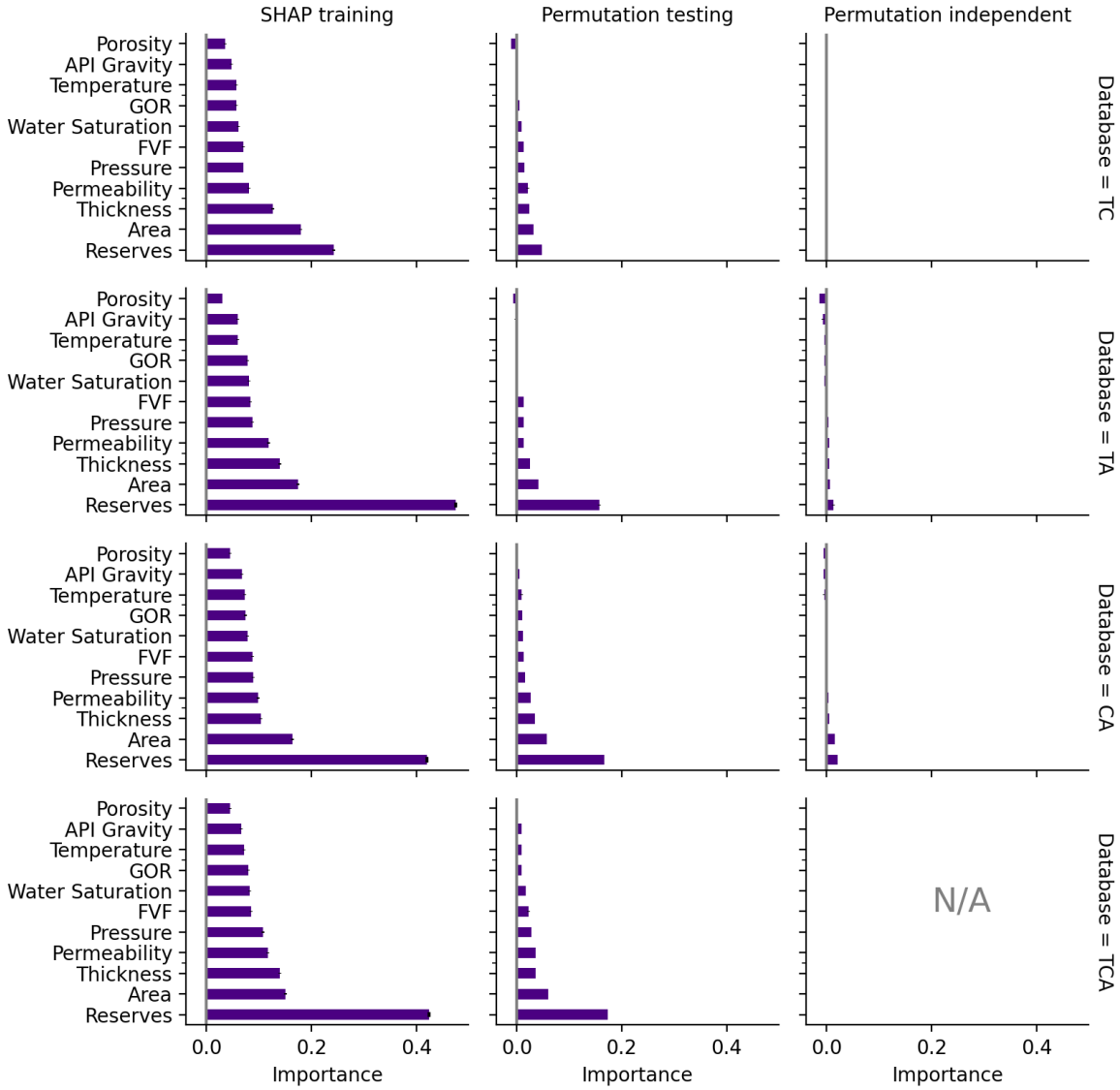

Figure 5: Feature importance for the XGBoost models for the TC (TORIS + commercial), TA (TORIS + Atlas, CA (commercial + Atlas), and TCA (TORIS + commercial & Atlas) databases for the training, testing, and independent datasets. The left row shows the average absolute SHAP value of each feature. The middle and right rows show the permutation feature importance. Uncertainty in the feature importance is visualized with black lines at the end of each bar. The independent dataset for TC is Atlas, for TA is the Commercial data, and for CA is TORIS.

 https://doi.org/10.69631/ipj.v2i3nr53

Permutation feature importance on the testing data reveals that several factors have low prediction performance. Porosity has a negative or zero performance in all testing datasets. For several datasets, API gravity and temperature were found to have negligible impact on model performance. On the other hand, reserves and area are consistently the two most important features. In the independent datasets, feature importance is highly degraded. The model trained on the TC database has no useful features when predicting recoveries from Atlas. Many features have zero or negative importance. Only reserves appear important for multiple datasets. Porosity has negative importance in two datasets.

4. DISCUSSION

The classification-based ML models constructed in this study showed satisfactory performances for the training and testing datasets. However, the performance of the models for the independent databases was poor. Dissimilarities in training and independent datasets are most probably why ML models predict an output feature unsatisfactorily for unseen data (61). Schaap and Leij (62) are among the first who demonstrated that regression-based ML models are database-dependent, meaning that they may not provide satisfactory estimations if evaluated using new and unseen data. As we stated earlier, ML models are not typically assessed using unseen data and independent databases. The database dependence of regression-based ML models in the estimation of oil and gas RFs was recently addressed by Roustazadeh et al. (61) who showed that the accuracy and reliability of regression-based ML models depended on data and size of databases used to train them. Those authors used the t-test method to statistically analyze the similarity in data. Although their results showed no significant differences between training and testing datasets (p-value >0.05), they reported p-values less than 0.05 by comparing input and output features in training datasets with those in independent databases, meaning that their trained datasets had significantly different values than their independent databases.

Another reason for the database dependence is the subjective way in which the features are measured. For instance, the thickness could be gross or net thickness, and the net thickness cutoff could be different between different datasets, and indeed between different interpreters. The average porosity could be averaged by well, by area, or by volume, and derived from core, wireline logs, or both. Water saturation could include clay-bound water or not, and it could be calculated from wireline logs, core measurements, water production, or a combination thereof. The dataset curators appear, from the results, to have been self-consistent in defining how features are measured, but that is unlikely to have been the case across all databases.

It is not straightforward to compare results of classification-based models with those of regression-based ones because evaluation parameters used in classification problems (e.g., accuracy and f1 score) are different from those used in regression cases (e.g., root mean square error and correlation coefficient). In the study by Roustazadeh et al. (61), the accuracy was quantified by root mean square error, and the coefficient of determination for independent databases was substantially less than that for training and testing datasets. In this work, although the performance of the classification-based ML models for the independent databases is not as satisfactory as for the training and testing datasets, the difference is not substantial, and the ML models still provide reasonable oil RF estimations, with over 50% within one class of the correct RF estimation (within-1 accuracy > 0.55 in **Table 3**).

Our feature importance results for the TC database disagree with those reported by Makhotin et al. (48) who applied the regression-based gradient boosting approach to estimate the oil RF. Those authors constructed two models based on pre- and post-production data using two databases i.e., TORIS, composed on 1381 oil reservoirs from the United States, and Proprietary, containing 1119 oil reservoirs from around the world collected by Russian oil companies. The results of their pre-production model showed that the top four features detected using the F-score were permeability, water saturation, viscosity, and API gravity (see Fig. 7 in 48). While we found permeability was among the top 4 factors, other factors were slightly less important. Although Makhotin et al. (48) found the reservoir thickness among the top five features, reserves was among the least impactful features in their pre-production

analysis. This disagreement could result from the feature importance conferred by their proprietary database, which we do not have access to.

In another study, using 18,000 oil reservoirs, Pooladi-Darvish et al. (60) applied the XGBoost model and trained it to estimate oil RF from water saturation, API gravity, pressure, porosity, permeability, temperature, viscosity, reservoir depth and thickness. Their results showed that only reservoir depth, permeability, porosity, oil viscosity, oil density and water saturation were the most important parameters. In our study, however, porosity was among the least influential parameters (see **Fig. 5**). This could be due to the large database that Pooladi-Darvish et al. (60) used in their study. The effect of the number of samples on ML models has been highlighted by Dawson et al. (15) and Ahmadisharaf et al. (3). Several studies demonstrated that by combining databases and increasing the number of samples, the accuracy of ML-based models improved (3, 21). Combining data from different resources, however, requires that the input and output parameters be consistently defined and determined among the databases. For instance, it is well documented that reserves has been defined differently in the literature (16, 17). One should also keep in mind that reservoirs from different databases are not necessarily at the same stage of production. Since RF changes with time, one should expect more accurate predictions when time of production is considered as an input feature.

All input and output features used in this study were numerical. We should point out that including some categorical data, such as reservoir rock origin (carbonate, sandstone, etc.) or reservoir conditions (e.g., fractured or not) may improve RF predictions. Therefore, further investigations are still required to predict RF from both numerical and categorical features.

5. CONCLUSION

In this study, we constructed classification-based ML models to estimate the oil RF from readily available data at the exploration stage. We collected thousands of reservoir observations by combining three databases. To address whether the classification-based ML models are database-dependent, we used four combinations of the three databases. The oil RF data were grouped into ten classes with 10% intervals. Using the XGBoost classification algorithm, we trained the ML models for each combination. Results showed that the constructed models were accurately trained for all the combinations except the one that had the lowest number of samples. The oil RF estimations for the testing datasets were within either the correct or a neighboring class 73-75% of the time. There was degradation between training and testing datasets. Models were tested on independent datasets to assess the database dependence of the ML models. Results showed that the accuracy of the models for the independent databases was less than that for the testing datasets. This could be due to inconsistencies in how the features were estimated. Further investigations are still required to improve the accuracy and reliability of ML models. The choice of ten classes in this study is also somewhat arbitrary, and it remains unclear how the results would differ if a narrower class interval (e.g., 5%) was used. Further research is needed to determine whether the accuracy of the XGBoost classification model improves or declines as the number of classes increases.

STATEMENTS AND DECLARATIONS

Supplementary Material

The supplementary material for this paper can be downloaded online here.

Acknowledgements

The authors are grateful to Amirhossein Yazdavar, Principal data scientist at Peerlogic, for the fruitful discussions and comments. Alireza Roustazadeh acknowledges the Department of Geology at Kansas State University for the financial support through the William J. Barret Fund for Excellence in Geology and the Gary & Kathie Sandlin Geology Scholarship. Behzad Ghanbarian acknowledges the University of Texas at Arlington for the support through the faculty start-up fund and STARs award. Frank Male was funded by startup grants from the Pennsylvania State University.

Author Contributions

Roustazadeh: Data curation, formal analysis, investigation, methodology, writing - original draft, Writing - review & editing. **Male:** Conceptualization, Resources, data curation, formal analysis, supervision, validation, visualization, investigation, methodology, writing - original draft, Writing - review & editing. **Ghanbarian:** Conceptualization, Resources, supervision, validation, methodology, visualization, funding acquisition, investigation, writing - original draft, Writing - review & editing. **Shadmand:** Conceptualization, Resources, supervision, validation, funding acquisition, investigation, writing - original draft. **Taslimitehrani:** Conceptualization, Resources, Supervision, Validation, investigation, writing - original draft, Methodology. **Lake:** Conceptualization, Resources, Supervision, Validation, Investigation, Methodology, Writing - original draft.

Conflicts of Interest

There are no conflicts of interest to declare.

Data, Code & Protocol Availability

The Python code that generated the models and figures as well as means, standard deviations, and distributions of input features used in this study are available at:

https://github.com/frank1010111/Estimating-RF-early-with-ML-classification/tree/frank-cleaning.

ORCID IDs

Alireza Roustazadeh https://orcid.org/0000-0002-3785-0522
Frank Male https://orcid.org/0000-0002-3402-5578
Behzad Ghanbarian https://orcid.org/0000-0002-7002-4193
Mohammad B. Shadmand https://orcid.org/0000-0002-3950-8640
Vahid Taslimitehrani https://orcid.org/0000-0003-2945-7496

REFERENCES

1. Agbadze, O. K., Qiang, C., & Jiaren, Y. (2022). Acoustic impedance and lithology-based reservoir porosity analysis using predictive machine learning algorithms. *Journal of Petroleum Science and Engineering*, 208, 109656. https://doi.org/10.1016/j.petrol.2021.109656
2. Ahmadi, M. A., & Chen, Z. (2019). Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs. *Petroleum*, 5(3), 271–284. https://doi.org/10.1016/j.petlm.2018.06.002
3. Ahmadisharaf, A., Nematirad, R., Sabouri, S., Pachepsky, Y., & Ghanbarian, B. (2024). Representative sample size for estimating saturated hydraulic conductivity via machine learning: A proof-of-concept study. *Water Resources Research*, 60(8), e2023WR036783. https://doi.org/10.1029/2023WR036783
4. Aliyuda, K., & Howell, J. (2019). Machine-learning algorithm for estimating oil-recovery factor using a combination of engineering and stratigraphic dependent parameters. *Interpretation*, 7(3), SE151–SE159. https://doi.org/10.1190/INT-2018-0211.1
5. Aliyuda, K., Howell, J., & Humphrey, E. (2020). Impact of geological variables in controlling oil-reservoir performance: An insight from a machine-learning technique. *SPE Reservoir Evaluation & Engineering*, 23(04), 1314–1327. https://doi.org/10.2118/201196-PA
6. Alpak, F. O., Araya–Polo, M., & Onyeagoro, K. (2019). Simplified dynamic modeling of faulted turbidite reservoirs: A deep-learning approach to recovery-factor forecasting for exploration. *SPE Reservoir Evaluation & Engineering*, 22(04), 1240–1255. https://doi.org/10.2118/197053-PA
7. Anifowose, F. A., Ewenla, A. O., & Eludiora, S. I. (2011). Prediction of oil and gas reservoir properties using support vector machines. *International Petroleum Technology Conference*, IPTC-14514-MS. https://doi.org/10.2523/IPTC-14514-MS
8. Burgess, G. L., Cross, K. K., & Kazanis, E. G. (2019, December 31). Outer Continental Shelf Estimated Oil and Gas Reserves Gulf of Mexico OCS Region. *U.S. Department of the Interior Bureau of Ocean Energy Management Gulf of Mexico OCS Region*. https://www.boem.gov/sites/default/files/documents/renewable-energy/state-activities/2019-EOGR.pdf
9. Carpenter, C. (2021). Machine-learning work flow identifies brittle, fracable, producible rock using drilling data. *Journal of Petroleum Technology*, 73(10), 61–62. https://doi.org/10.2118/1021-0061-JPT

10. Chen, H., Chen, H., Liu, Z., Sun, X., & Zhou, R. (2020). Analysis of factors affecting the severity of automated vehicle crashes using XGBoost model combining POI data. *Journal of Advanced Transportation*, 2020, 1–12. https://doi.org/10.1155/2020/8881545
11. Chen, L., Gao, X., & Li, X. (2021). Using the motor power and XGBoost to diagnose working states of a sucker rod pump. *Journal of Petroleum Science and Engineering*, 199, 108329. https://doi.org/10.1016/j.petrol.2020.108329
12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
13. Chen, Z., Yu, W., Liang, J.-T., Wang, S., & Liang, H. (2022). Application of statistical machine learning clustering algorithms to improve EUR predictions using decline curve analysis in shale-gas reservoirs. *Journal of Petroleum Science and Engineering*, 208, 109216. https://doi.org/10.1016/j.petrol.2021.109216
14. Cronshaw, M. (2021). Energy in Perspective (p. 222). Springer International Publishing. https://doi.org/10.1007/978-3-030-63541-1
15. Dawson, H. L., Dubrule, O., & John, C. M. (2023). Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification. *Computers & Geosciences*, 171, 105284. https://doi.org/10.1016/j.cageo.2022.105284
16. Denney, D. (2007). Reserves and resources classification, definitions, and guidelines: Defining the standard! *Journal of Petroleum Technology*, 59(12), 63–67. https://doi.org/10.2118/1207-0063-JPT
17. Desorcy, G. J., Warne, G. A., Ashton, B. R., Campbell, G. R., Collyer, D. R., et al. (1993). Definitions and guidelines for classification of oil and gas reserves. *Journal of Canadian Petroleum Technology*, 32(05). https://doi.org/10.2118/93-05-01
18. Dias, L. O., Bom, C. R., Faria, E. L., Valentín, M. B., Correia, M. D., et al. (2020). Automatic detection of fractures and breakouts patterns in acoustic borehole image logs using fast-region convolutional neural networks. *Journal of Petroleum Science and Engineering*, 191, 107099. https://doi.org/10.1016/j.petrol.2020.107099
19. Dong, Y., Qiu, L., Lu, C., Song, L., Ding, Z., Yu, Y., & Chen, G. (2022). A data-driven model for predicting initial productivity of offshore directional well based on the physical constrained eXtreme gradient boosting (XGBoost) trees. *Journal of Petroleum Science and Engineering*, 211, 110176. https://doi.org/10.1016/j.petrol.2022.110176
20. Esfandi, T., Sadeghnejad, S., & Jafari, A. (2024). Effect of reservoir heterogeneity on well placement prediction in CO_2-EOR projects using machine learning surrogate models: Benchmarking of boosting-based algorithms. *Geoenergy Science and Engineering*, 233, 212564. https://doi.org/10.1016/j.geoen.2023.212564
21. Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, 58(4), e2021WR029583. https://doi.org/10.1029/2021WR029583
22. Fetkovich, M. J., Fetkovich, E. J., & Fetkovich, M. D. (1996). Useful concepts for decline-curve forecasting, reserve estimation, and analysis. *SPE Reservoir Engineering*, 11(01), 13–22. https://doi.org/10.2118/28628-PA
23. Gu, Y., Zhang, D., Lin, Y., Ruan, J., & Bao, Z. (2021). Data-driven lithology prediction for tight sandstone reservoirs based on new ensemble learning of conventional logs: A demonstration of a Yanchang member, Ordos Basin. *Journal of Petroleum Science and Engineering*, 207, 109292. https://doi.org/10.1016/j.petrol.2021.109292
24. Gupta, S., Saputelli, L. A., Verde, A., Vivas, J. A., & Narahara, G. M. (2016). Application of an advanced data analytics methodology to predict hydrocarbon recovery factor variance between early phases of appraisal and post-sanction in gulf of mexico deep offshore assets. *Offshore Technology Conference*, D041S056R005. https://doi.org/10.4043/27127-MS
25. Han, B., & Bian, X. (2018). A hybrid PSO-SVM-based model for determination of oil recovery factor in the low-permeability reservoir. *Petroleum*, 4(1), 43–49. https://doi.org/10.1016/j.petlm.2017.06.001
26. Hartmann, D. J., & Beaumont, E. A. (1999). Predicting reservoir system quality and performance. In E. A. Beaumont & N. H. Foster, *Exploring for Oil and Gas Traps*. American Association of Petroleum Geologists. https://doi.org/10.1306/TrHbk624C9
27. He, M., Gu, H., & Xue, J. (2022). Log interpretation for lithofacies classification with a robust learning model using stacked generalization. *Journal of Petroleum Science and Engineering*, 214, 110541. https://doi.org/10.1016/j.petrol.2022.110541
28. Helmy, T., & Fatai, A. (2010). Hybrid computational intelligence models for porosity and permeability prediction of petroleum reservoirs. *International Journal of Computational Intelligence and Applications*, 09(04), 313–337. https://doi.org/10.1142/S1469026810002902
29. Holdaway, K. R. (Ed.). (2014). Harness Oil and Gas Big Data with Analytics (1st ed.). Wiley. https://doi.org/10.1002/9781118910948

30. Kaczmarczyk, R., Herbas, J., & Del Castillo, J. (2013). Approximations of primary, secondary and tertiary recovery factor in viscous and heavy oil reservoirs. *SPE Offshore Europe Oil and Gas Conference and Exhibition*, SPE-166583-MS. https://doi.org/10.2118/166583-MS
31. Kalam, S., Yousuf, U., Abu-Khamsin, S. A., Waheed, U. B., & Khan, R. A. (2022). An ANN model to predict oil recovery from a 5-spot waterflood of a heterogeneous reservoir. *Journal of Petroleum Science and Engineering*, 210, 110012. https://doi.org/10.1016/j.petrol.2021.110012
32. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. https://doi.org/10.1016/j.patter.2023.100804
33. Karacan, C. Ö. (2020). A fuzzy logic approach for estimating recovery factors of miscible CO_2-EOR projects in the United States. *Journal of Petroleum Science and Engineering*, 184, 106533. https://doi.org/10.1016/j.petrol.2019.106533
34. Knopp, C. R., & Ramsey, L. A. (1960). Correlation of oil formation volume factor and solution gas-oil ratio. *Journal of Petroleum Technology*, 12(08), 27–29. https://doi.org/10.2118/1433-G
35. Kong, B., Chen, Z., Chen, S., & Qin, T. (2021). Machine learning-assisted production data analysis in liquid-rich Duvernay Formation. *Journal of Petroleum Science and Engineering*, 200, 108377. https://doi.org/10.1016/j.petrol.2021.108377
36. Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, 39(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4
37. Kumar, M., Swaminathan, K., Rusli, A., & Thomas-Hy, A. (2022). Applying data analytics & machine learning methods for recovery factor prediction and uncertainty modelling. *SPE Asia Pacific Oil & Gas Conference and Exhibition*, D021S008R003. https://doi.org/10.2118/210769-MS
38. Lake, L., Johns, R. T., Rossen, W. R., & Pope, G. A. (2014). Fundamentals of enhanced oil recovery. *Society of Petroleum Engineers*. https://doi.org/10.2118/9781613993286
39. Lee, B. B., & Lake, L. W. (2015). Using data analytics to analyze reservoir databases. *SPE Annual Technical Conference and Exhibition*, D031S030R008. https://doi.org/10.2118/174900-MS
40. Li, S., Zhou, K., Zhao, L., Xu, Q., & Liu, J. (2022). An improved lithology identification approach based on representation enhancement by logging feature decomposition, selection and transformation. *Journal of Petroleum Science and Engineering*, 209, 109842. https://doi.org/10.1016/j.petrol.2021.109842
41. Lin, J., De Weck, O., & MacGowan, D. (2012). Modeling epistemic subsurface reservoir uncertainty using a reverse Wiener jump–diffusion process. *Journal of Petroleum Science and Engineering*, 84–85, 8–19. https://doi.org/10.1016/j.petrol.2012.01.015
42. Lin, W.-C., & Tsai, C.-F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4
43. Ling, K., Wu, X., Zhang, H., & He, J. (2013). Tactics and pitfalls in production decline curve analysis. *SPE Production and Operations Symposium*, SPE-164503-MS. https://doi.org/10.2118/164503-MS
44. Liu, W., Liu, W. D., Gu, J., & Shen, X. (2019). Predictive model for water absorption in sublayers using a machine learning method. *Journal of Petroleum Science and Engineering*, 182, 106367. https://doi.org/10.1016/j.petrol.2019.106367
45. Long, R. (2016). TORIS: An Integrated Decision Support System for Petroleum E&P Policy Evaluation. USGEO, AmeriGEO. https://data.amerigeoss.org/dataset/toris-an-integrated-decision-support-system-for-petroleum-e-p-policy-evaluation
46. Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions (No. arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874
47. Mahmoud, A., Elkatatny, S., Chen, W., & Abdulraheem, A. (2019). Estimation of oil recovery factor for water drive sandy reservoirs through applications of artificial intelligence. *Energies*, 12(19), 3671. https://doi.org/10.3390/en12193671
48. Makhotin, I., Orlov, D., Koroteev, D., Burnaev, E., Karapetyan, A., & Antonenko, D. (2022). Machine learning for recovery factor estimation of an oil reservoir: A tool for derisking at a hydrocarbon asset evaluation. Petroleum, 8(2), 278–290. https://doi.org/10.1016/j.petlm.2021.11.005
49. Male, F., Jensen, J. L., & Lake, L. W. (2020). Comparison of permeability predictions on cemented sandstones with physics-based and machine learning approaches. *Journal of Natural Gas Science and Engineering*, 77, 103244. https://doi.org/10.1016/j.jngse.2020.103244
50. Maselugbo, A. O., Onolemhemhen, R. U., Denloye, A. O., Salufu, S. O., & Isehunwa, S. O. (2017). Optimization of gas recovery using co-production technique in water drive reservoir. *Journal of Petroleum and Gas Engineering*, 8(6), 42–48. https://doi.org/10.5897/JPGE2017.0269
51. Matkerim, B., Mukhanbet, A., Kassymbek, N., Daribayev, B., Mustafin, M., & Imankulov, T. (2024). Machine learning analysis using the black oil model and parallel algorithms in oil recovery forecasting. *Algorithms*, 17(8), 354. https://doi.org/10.3390/a17080354

52. Meng, H., Wang, X., & Wang, X. (2018). Expressway crash prediction based on traffic big data. *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 11–16. https://doi.org/10.1145/3297067.3297093
53. Mohaghegh, S. (2000). Virtual-intelligence applications in petroleum engineering: Part 1—artificial neural networks. *Journal of Petroleum Technology*, 52(9). https://doi.org/10.2118/58046-MS
54. Mousavi, S. M., Bakhtiarimanesh, P., Enzmann, F., Kersten, M., & Sadeghnejad, S. (2024). Machine-learned surrogate models for efficient oil well placement under operational reservoir constraints. *SPE Journal*, 29(01), 518–537. https://doi.org/10.2118/217467-PA
55. Omoniyi, O. A., & Adeolu, S. (2014). Decline Curve Analysis and Material Balance, as Methods for Estimating Reserves (A Case Study of D4 and E1 Fields). *The International Journal of Innovative Research and Development*, 3(11), 207–218. https://www.internationaljournalcorner.com/index.php/ijird_ojs/article/view/135465
56. Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, 208, 109520. https://doi.org/10.1016/j.petrol.2021.109520
57. Parish, R. G., Calderbank, V. J., Watkins, A. J., Muggeridge, A. H., Goode, A. T., & Robinson, P. R. (1993). Effective history matching: The application of advanced software techniques to the history-matching process. *SPE Symposium on Reservoir Simulation*, SPE-25250-MS. https://doi.org/10.2118/25250-MS
58. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al. (2011). Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research, 12, 2825–2830.
59. Pirizadeh, M., Alemohammad, N., Manthouri, M., & Pirizadeh, M. (2021). A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods. *Journal of Petroleum Science and Engineering*, 198, 108214. https://doi.org/10.1016/j.petrol.2020.108214
60. Pooladi-Darvish, M., Tabatabaie, S. H., & Rodriguez Cadena, C. (2022). Development of a machine learning technique in conjunction with reservoir complexity index to predict recovery factor using data from 18,000 reservoirs. *ADIPEC*, D032S173R006. https://doi.org/10.2118/211410-MS
61. Roustazadeh, A., Ghanbarian, B., Shadmand, M. B., Taslimitehrani, V., & Lake, L. W. (2024). Estimating hydrocarbon recovery factor at reservoir scale via machine learning: Database-dependent accuracy and reliability. *Engineering Applications of Artificial Intelligence*, 128, 107500. https://doi.org/10.1016/j.engappai.2023.107500
62. Schaap, M. G., & Leij, F. J. (1998). Database-related accuracy and uncertainty of pedotransfer functions. *Soil Science*, 163(10), 765–779. https://doi.org/10.1097/00010694-199810000-00001
63. Shapley, L. S. (1953). 17. A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games* (AM-28), Volume II (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018
64. Sharma, A., Srinivasan, S., & Lake, L. W. (2010). Classification of oil and gas reservoirs based on recovery factor: A data-mining approach. *SPE Annual Technical Conference and Exhibition*, SPE-130257-MS. https://doi.org/10.2118/130257-MS
65. Sheng, J. J. (2013). Surfactant enhanced oil recovery in carbonate reservoirs. In *Enhanced Oil Recovery Field Case Studies* (pp. 281–299). Elsevier. https://doi.org/10.1016/B978-0-12-386545-8.00012-9
66. Srivastava, P., Wu, X., Amirlatifi, A., & Devegowda, D. (2016). Recovery factor prediction for deepwater gulf of Mexico oilfields by integration of dimensionless numbers with data mining techniques. *SPE Intelligent Energy International Conference and Exhibition*, SPE-181024-MS. https://doi.org/10.2118/181024-MS
67. Talluru, G., & Wu, X. (2017). Using data analytics on dimensionless numbers to predict the ultimate recovery factors for different drive mechanisms of gulf of Mexico oil fields. *SPE Annual Technical Conference and Exhibition*, D031S030R008. https://doi.org/10.2118/187269-MS
68. Tang, J., Fan, B., Xiao, L., Tian, S., Zhang, F., et al. (2021). A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. *SPE Journal*, 26(01), 482–497. https://doi.org/10.2118/204224-PA
69. Tewari, S., Dwivedi, U. D., & Shiblee, M. (2019). Assessment of big data analytics based ensemble estimator module for the real-time prediction of reservoir recovery factor. *SPE Middle East Oil and Gas Show and Conference*, D041S038R003. https://doi.org/10.2118/194996-MS
70. Tunkiel, A. T., Sui, D., & Wiktorski, T. (2022). Impact of data pre-processing techniques on recurrent neural network performance in context of real-time drilling logs in an automated prediction framework. *Journal of Petroleum Science and Engineering*, 208, 109760. https://doi.org/10.1016/j.petrol.2021.109760
71. Vo Thanh, H., Sheini Dashtgoli, D., Zhang, H., & Min, B. (2023). Machine-learning-based prediction of oil recovery factor for experimental CO_2-Foam chemical EOR: Implications for carbon utilization projects. *Energy*, 278, 127860. https://doi.org/10.1016/j.energy.2023.127860
72. Woods, R. W., & Muskat, M. (1945). An analysis of material-balance calculations. *Transactions of the AIME*, 160(01), 124–139. https://doi.org/10.2118/945124-G

73. Zhao, X., Chen, X., Huang, Q., Lan, Z., Wang, X., & Yao, G. (2022). Logging-data-driven permeability prediction in low-permeable sandstones based on machine learning with pattern visualization: A case study in Wenchang A Sag, Pearl River Mouth Basin. *Journal of Petroleum Science and Engineering*, 214, 110517. https://doi.org/10.1016/j.petrol.2022.110517