RESEARCH ARTICLE

Application of Machine Learning for Effective Screening of Enhanced Oil Recovery Methods

Jawad Ali[1,*], **Ubedullah Ansari**[1], **Fateh Ali**[1], **Tariq Javed**[1] **and Imran Ahmed Hullio**[1]

[1] Institute of Petroleum and Natural Gas Engineering, Mehran University of Engineering and Technology, Jamshoro 76062, Pakistan

Abstract

Selecting the most suitable enhanced oil recovery (EOR) technique remains challenging due to severe class imbalance in historical datasets and the limitations of traditional screening criteria. To address data imbalance while preserving domain knowledge, this study proposes a novel machine learning framework that incorporates domain-informed synthetic data generation strictly constrained by established EOR screening criteria. An initial dataset of 583 documented EOR projects was compiled from field reports and public databases. After rigorous cleaning, 575 valid samples were retained and subsequently augmented to 760 balanced instances (class sizes ranging from 60–110 samples per class). This reduced the imbalance ratio from 123:1 to approximately 1.8:1. The augmented dataset was processed using principal component analysis (PCA) for dimensionality reduction, followed by hyperparameter tuning and 5-fold cross-validation. Among the evaluated models, K-Nearest Neighbors (KNN) and Random Forest achieved the highest macro-averaged performance (F1-score of 0.89 and 0.85, respectively). The results demonstrate that domain-guided synthetic data generation significantly improves model accuracy and robustness for multi-class EOR screening, offering reservoir engineers a reliable, machine learning-supported decision-making tool.

Keywords: EOR screening, machine learning, screening criteria, imbalanced data, multi-class classification, enhanced oil recovery.

1 Introduction

Enhanced oil recovery (EOR) represents a pivotal area within reservoir engineering, dedicated to recovering residual oil from mature reservoirs, as illustrated in Figure 1. The selection of an optimal EOR method poses a persistent challenge, owing to the complex interplay of reservoir and fluid properties—such as depth, temperature, porosity, permeability, and oil characteristics [1]. Although traditional screening guidelines have seen extensive use, their effectiveness is often constrained by limited historical data and reliance on subjective expert interpretation.

Despite recent advances in applying machine learning to EOR screening [2, 3], existing approaches remain severely limited by two primary challenges: severe class imbalance in historical datasets and the lack of

Submitted: 27 November 2025
Accepted: 10 February 2026
Published: 27 February 2026

Vol. 2, **No.** 1, 2026.
10.62762/RS.2025.333184

***Corresponding author:**
✉ Jawad Ali
jawadaliv28@gmail.com

Citation
Ali, J., Ansari, U., Ali, F., Javed, T., & Hullio, I. A. (2026). Application of Machine Learning for Effective Screening of Enhanced Oil Recovery Methods. *Reservoir Science*, 2(1), 65–80.

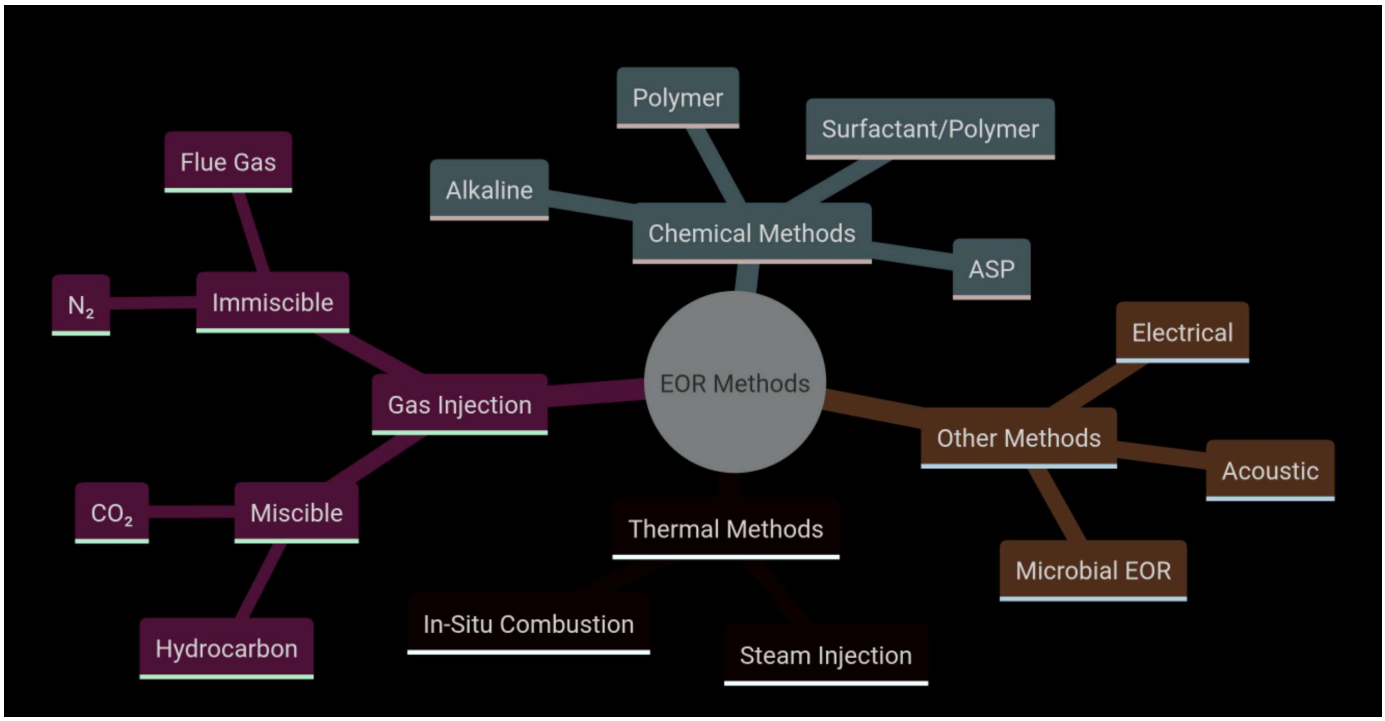


Figure 1. Classification of enhanced oil recovery (EOR) methods.

domain-informed synthetic data generation. As a result, as illustrated in Figure 4, models tend to exhibit strong bias toward the majority class while delivering poor predictive performance on underrepresented classes.

To address these limitations, this study proposes a novel machine learning framework that incorporates domain-informed synthetic data generation, strictly guided by established EOR screening criteria, while systematically comparing multiple algorithms [4]. The approach is built upon an initial compilation of 583 documented EOR field projects. Following rigorous data cleaning (detailed in Section 2), 575 high-quality samples were retained for subsequent modeling. Given the severe class imbalance—some EOR techniques were represented by very few examples—the dataset was augmented through controlled oversampling and synthetic sample generation, increasing the total to approximately 760 balanced instances and reducing the imbalance ratio from 123:1 to roughly 1.8:1. Model performance was further enhanced via grid search hyperparameter optimization applied to four selected algorithms, followed by rigorous evaluation using 5-fold cross-validation, which yields more reliable performance estimates [5].

There are three contributions to this research work:

1. Constructing a carefully curated dataset with relevant domain insights.
2. Comparing four machine learning algorithms to find the best one for selecting EOR techniques.
3. Introducing a practical tool to help choose most suitable EOR method for a reservoir.

By combining domain knowledge with advanced machine learning techniques, this research work provides a more robust and generalizable approach to EOR screening, providing both methodological and practical assistance to reservoir engineering.

2 Methodology

2.1 EOR Data Source and Screening Criteria

The overall workflow is illustrated in the flowchart presented in Figure 2. The dataset for this study was compiled from diverse sources, including field reports, peer-reviewed literature, and publicly available EOR databases. A total of 583 documented EOR projects were collected, encompassing a wide range of reservoir types and geographical regions [6–8]. These projects provide comprehensive details on fluid properties (e.g., viscosity and API gravity), reservoir characteristics (e.g., temperature, porosity, permeability, and depth), and the EOR techniques employed.

The dataset construction was guided by established EOR screening criteria, which define the optimal

Table 1. Screening criteria for enhanced oil recovery (EOR) methods based on fluid and reservoir properties.

Parameter	Nitrogen and Flue Gas	Hydrocarbon	Carbon Dioxide (CO_2)	Immiscible Gases	Micellar / Polymer ASP & Alkaline Flooding	Polymer Flooding	Combustion	Steam
Oil API Gravity	35 (avg. 48)	23 (avg. 41)	22 (avg. 36)	12	20 (avg. 35)	15–40	10 (avg. 16)	8–13.5 (avg. 13.5)
Oil Viscosity (cp)	0.4 (avg. 0.2)	3 (avg. 0.5)	10 (avg. 1.5)	600	35 (avg. 13)	10–150	5000 (avg. 1200)	200,000 (avg. 4700)
Composition	High % C_1–C_7	High % C_2–C_7	High % C_5–C_{12}	Not critical	Light/intermediate; some organic acids (for alkaline floods)	Not critical	Some asphaltic components	Not critical
Oil Saturation (PV Fraction)	0.40 (avg. 0.75)	0.30 (avg. 0.80)	0.20 (avg. 0.55)	0.35 (avg. 0.70)	0.35 (avg. 0.53)	0.70 (avg. 0.80)	0.50 (avg. 0.72)	0.40 (avg. 0.66)
Formation Type	Sandstone or Carbonate	Sandstone or Carbonate	Sandstone or Carbonate	Not critical	Sandstone preferred	Sandstone preferred	High-porosity sandstone	High-porosity sandstone
Net Thickness (ft)	Thin unless dipping	Thin unless dipping	Wide range	Not critical if dipping	Not critical	Not critical	10 ft	20 ft
Average Permeability (md)	Not critical	Not critical	Not critical	Not critical	10 md (avg. 450 md)	10 md (avg. 800 md)	≥50	200
Depth (ft)	6000	4000	2500	1800	9000 (avg. 3250)	9000	11500 (avg. 3500)	4500
Temperature (°F)	Not critical	Not critical	Not critical	Not critical	200	200	≥100	Not critical

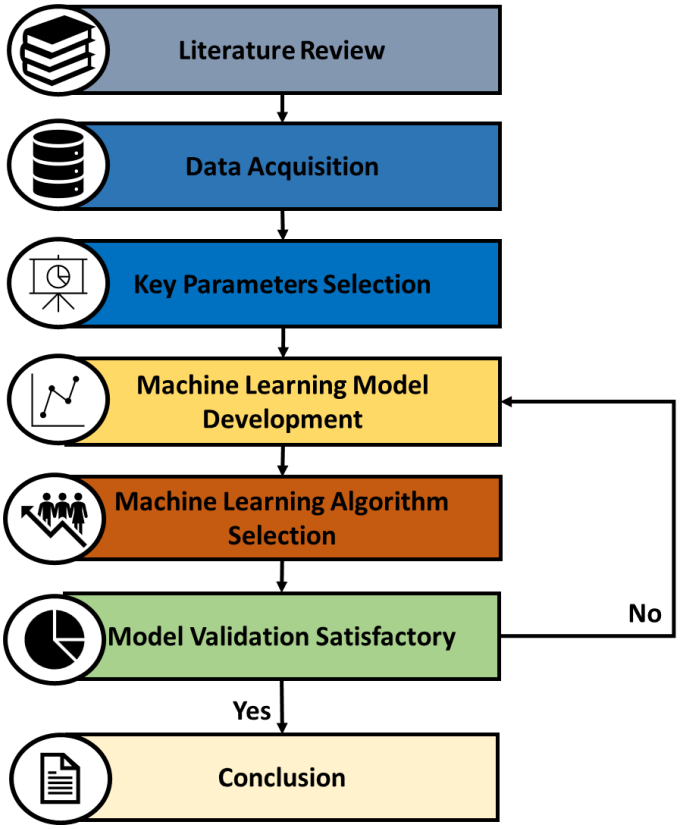


Figure 2. Research Methodology flow chart.

ranges of fluid and reservoir properties that enhance the likelihood of success for specific EOR techniques. For instance, polymer flooding is generally more suitable for shallower reservoirs, whereas miscible gas injection methods are typically associated with deeper reservoirs capable of maintaining higher pressures [6, 8]. By adhering to these predefined ranges, the compiled dataset ensured consistency with domain knowledge while incorporating real-world field cases.

The incorporation of screening criteria during the data collection stage is particularly critical, as it enables the generation of synthetic data that remains geologically realistic (see Table 1). This approach ensures that subsequent augmentation steps—such as oversampling and balancing—preserve geological validity rather than relying solely on statistical considerations.

2.2 Data Collection

The research data were collected from a variety of sources, encompassing a broad spectrum of successful Enhanced Oil Recovery (EOR) projects across different geographic regions and reservoir conditions. The dataset provides comprehensive information on EOR projects worldwide. Initially, it comprised 583 EOR projects, representing diverse characteristics such as fluid properties, geological formations, historical EOR techniques, and their corresponding performance outcomes. The data were obtained from reputable energy organizations, industry reports, peer-reviewed publications, and publicly available databases.

2.3 Data Cleaning

To guarantee the integrity and correctness of the gathered dataset, a thorough data cleaning procedure was implemented. This entailed locating and resolving any discrepancies, outliers, or inconsistencies in the data.

2.4 Missing Value Treatment

We identified missing values across the dataset and handled them using domain-informed imputation strategies rather than simple deletion, which would

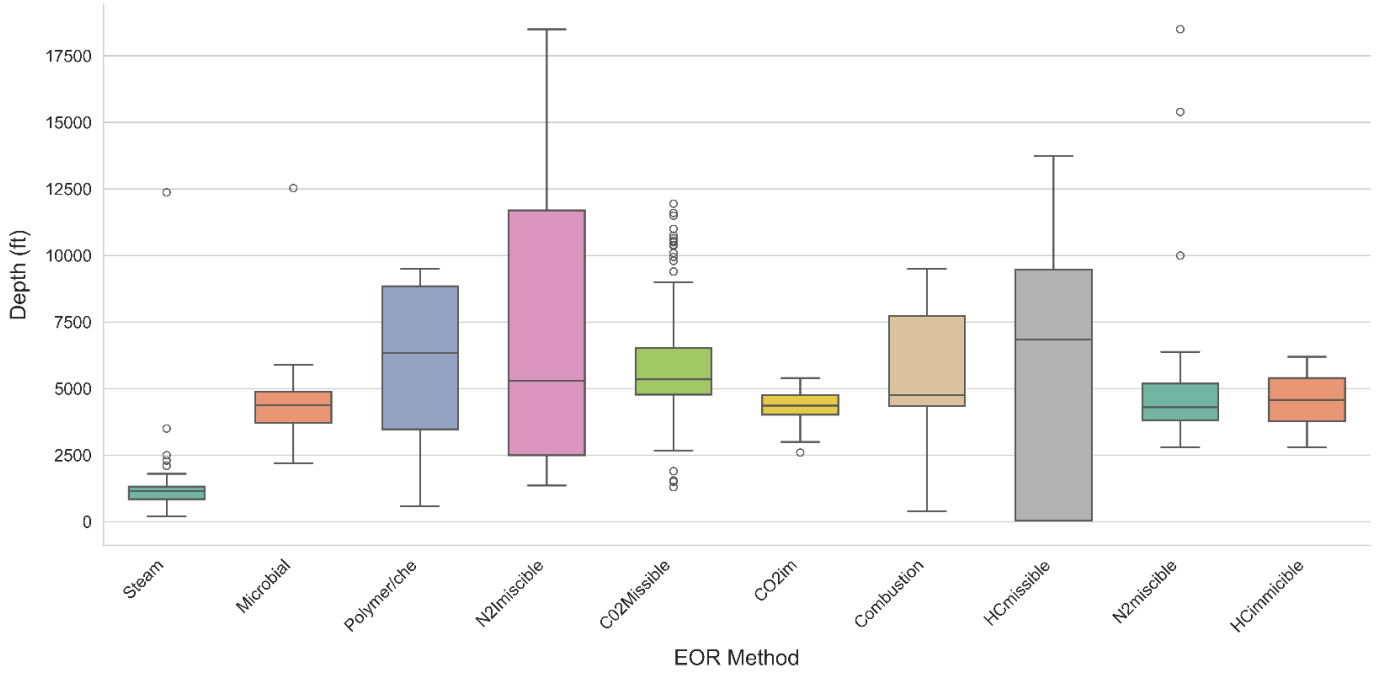


Figure 3. Box plot (EOR methods vs Depth).

have reduced our already limited sample size. For continuous variables (depth, viscosity, API gravity, temperature, permeability, oil saturation), we employed median imputation within each EOR class. This approach preserves the central tendency of each EOR method's characteristic parameter ranges, ensuring that imputed values align with the screening criteria for that specific method. For example, missing depth values in CO_2 miscible samples were replaced with the median depth of all CO_2 miscible projects, ensuring the imputed value respected the >2500 ft screening criterion.

For categorical variables such as formation type, mode imputation was employed: missing values were replaced with the most frequent category within the corresponding EOR class. For example, if a polymer flooding sample lacked formation type information, "Sandstone" was imputed, consistent with the preferred formation specified for polymer methods in the screening criteria (Table 1). Samples with more than 30% of features missing were excluded from the dataset, as excessive imputation for a single record would introduce substantial uncertainty and compromise reliability. This threshold was chosen to retain only those samples with sufficient original data to support credible analysis.

2.5 Outlier Detection and Treatment

Boxplots were used to detect outliers in continuous variables across each EOR class, as illustrated in Figure 3. Outliers were defined as values falling beyond 1.5 times the interquartile range (IQR) from the first and third quartiles. Rather than automatically removing all statistically identified outliers, domain knowledge was applied to differentiate between true anomalies and physically valid extreme cases.

For instance, Figure 3 reveals that the N_2 miscible class includes a maximum depth of 18,000 feet, while the majority of samples cluster between 3,000 and 4,000 feet. Although this extreme value qualifies as a statistical outlier, it corresponds to a documented, real-world deep reservoir case. Such outliers were retained if they remained within the physically feasible ranges specified by the EOR screening criteria in Table 1 and could be corroborated by published literature.

Conversely, we removed clear data entry errors such as negative viscosity values, API gravity exceeding 100°, or temperatures below surface conditions as these violate physical constraints. We also removed samples where parameter combinations were geologically implausible, such as very shallow reservoirs (<500 ft) with extremely high temperatures (>300°F), which would violate normal geothermal gradients.

2.6 Data Validation

Following imputation and outlier treatment, all retained samples were validated to ensure they fell within the acceptable ranges specified by the EOR screening criteria for their respective classes (Table 1).

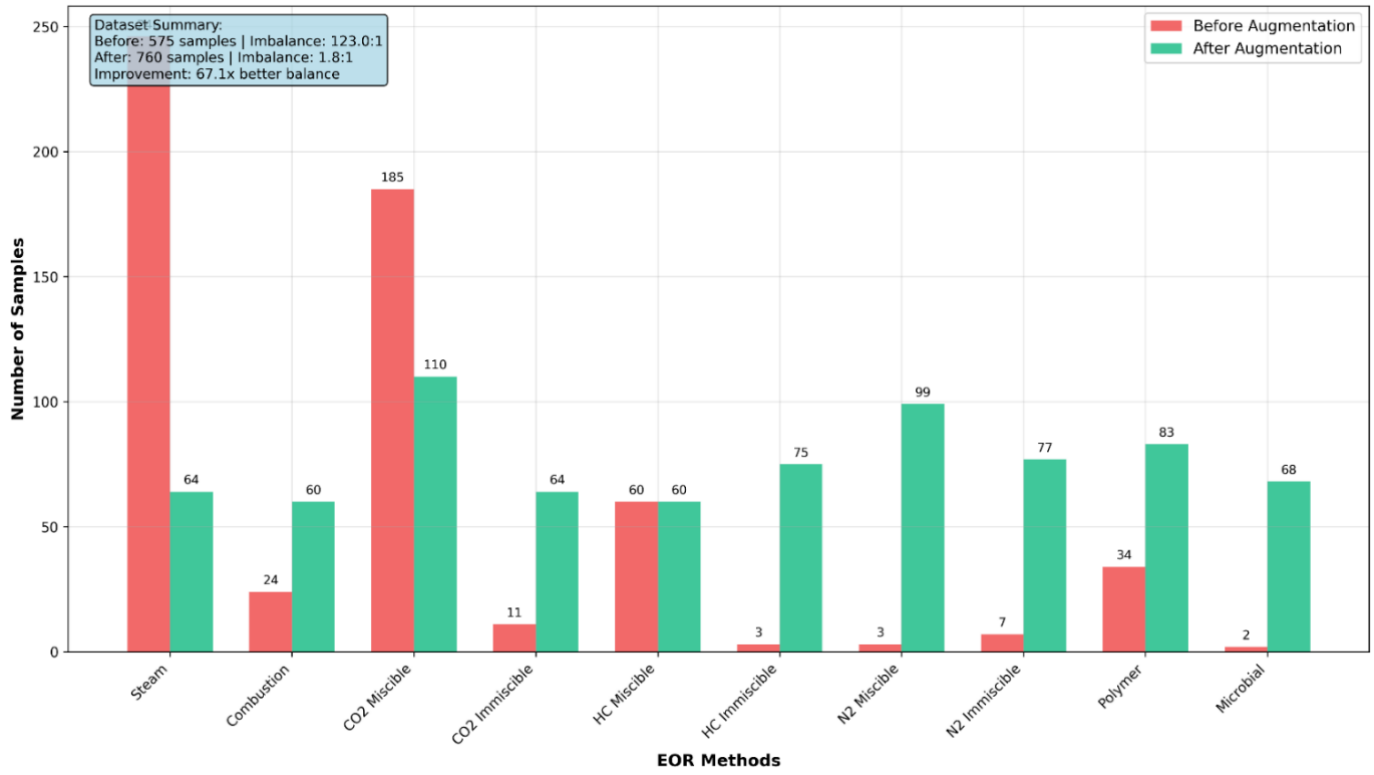


Figure 4. Sample Count Comparison Before and After Augmentation (Before: 575 samples with 123:1 imbalance; After: 760 samples with 1.8:1 imbalance).

This step confirmed that the cleaned dataset preserved geological realism and domain consistency. Overall, the data cleaning process reduced the original 583 samples to 575 high-quality samples, with only 8 samples (1.4%) excluded due to excessive missing values or evident data quality issues.

2.7 Data Transformation

Following data cleaning, the dataset was prepared for modeling through standard transformation techniques. Categorical variables were encoded using appropriate methods, such as one-hot encoding, to convert them into a format suitable for machine learning algorithms. Numerical features were scaled or normalized to ensure uniform ranges across all variables, thereby facilitating consistent model training and performance.

2.8 Synthetic Data Generation and Oversample

To address the severe class imbalance in the dataset [10], this study introduces a novel, domain-informed synthetic data augmentation approach that integrates oversampling with controlled synthetic sample generation, as illustrated in Figure 4 [9]. Unlike conventional statistical resampling techniques, our method is strictly constrained by established EOR screening criteria, ensuring that all generated instances remain geologically realistic and representative of valid reservoir conditions.

Through this combined strategy, the cleaned dataset of 575 samples was expanded to 760 instances. Although perfect balance was not achieved (with class sizes ranging from 60 to 110 samples, averaging approximately 108 per class), the approach substantially improved equity across classes and reduced the imbalance ratio from 123:1 to approximately 1.8:1. By mitigating bias toward majority classes, this domain-guided technique enabled more equitable learning across all EOR methods.

2.8.1 How Synthetic Data was Generated

Our method is straightforward: we generate new samples by creating variations of existing samples, but we strictly keep all parameters within the ranges specified in Table 1 for each EOR method.

Generation Process:

1. Identify the target EOR class that needs more samples (e.g., CO_2 Miscible has only 11 samples).
2. Extract the valid ranges from Table 1 for that EOR method. For CO_2: API Gravity >22°, Depth >2500 ft, Viscosity <10 cp, Oil Saturation >0.20.

3. Select a real sample from that class as a starting point.
4. Generate new values by creating small variations around the seed sample, but always staying within the Table 1 ranges.
    - If seed has API=30°, we might generate API=32° or API=28°, but never below 22° (Table 1 minimum)
    - If seed has Depth=3000 ft, we might generate Depth=3500 ft or Depth=2800 ft, but never below 2500 ft (Table 1 minimum)
5. Check all parameters to ensure they fall within the screening criteria ranges for that EOR class.
6. Accept the sample if all parameters are within valid ranges; otherwise, generate again.

Generating CO_2 Miscible Samples:

Let's say we need to generate more CO_2 samples. Table 1 tells us CO_2 works when:

- API Gravity: >22° (average 36°)
- Depth: >2500 ft
- Viscosity: <10 cp (average 1.5 cp)
- Oil Saturation: >0.20 (average 0.55)

Starting with a real CO_2 sample:

- API = 30°
- Depth = 3000 ft
- Viscosity = 3 cp
- Saturation = 0.50

We have a new sample:

- API = 32.5° (slightly higher, but still >22°)
- Depth = 3350 ft (slightly deeper, still >2500 ft)
- Viscosity = 2.8 cp (slightly lower, still <10 cp)
- Saturation = 0.53 (slightly higher, still >0.20)

All parameters respect Table 1 ranges, so this is a valid synthetic CO_2 sample.

2.9 Why This Method is Better Than Traditional Approaches

Traditional SMOTE (most common method):

Most data generation methods interpolate samples by averaging features but ignores EOR screening criteria, producing invalid data like CO_2 injection at 2000 ft depth (violating >2500 ft minimum), rendering outputs unrealistic for petroleum engineering.

DI-SDG Method:

This method uses the screening criteria in Table 1 as boundaries, generating samples only within proven ranges, where EOR methods actually work. And cannot create impossible combinations because we enforce range limits which result in synthetic data that represents realistic reservoir conditions. The key Difference such that SMOTE just does math without understanding petroleum engineering. This method incorporates field knowledge from Table 1, ensuring every synthetic sample could actually exist in a real reservoir.

2.10 Data Splitting

Following oversampling and synthetic data generation, the dataset was partitioned into training and testing sets. Approximately 80% of the samples (608 instances) were allocated to the training set, while the remaining 20% (152 samples) formed the test set. This split enabled a robust evaluation of model performance on unseen data, ensuring reliable assessment of accuracy and generalization. Data splitting can be viewed as a statistical sampling problem [11]. Accordingly, various traditional statistical sampling techniques can be employed to divide the data [12]. Based on their underlying principles, objectives, and computational complexity, these methods may be categorized as follows:

- Simple random sampling (SRS)
- Trial-and-error methods
- Systematic sampling
- Convenience sampling
- CADEX and DUPLEX
- Stratified sampling

Despite their widespread use and simplicity, the methods (such as SRS and trial-and-error methods) suffer from high model performance variance. Additional methods are efficient and deterministic, but they are limited to particular kinds of datasets (e.g. convenience and systematic sampling). At the expense of increased computational costs, the more complex techniques (such as CADEX, DUPLEX, and stratified sampling) take advantage of the data's structure to obtain reliable results.

Table 2. Rationale for the selected models.

Model	Key Strengths	EOR Relevance
Decision Tree	Interpretable hierarchy; no preprocessing needed for mixed features.	Mimics engineer screening trees (e.g., API gravity thresholds).
Random Forest	Ensemble robustness to imbalance; feature importance.	Ranks parameters like viscosity, aiding reservoir prioritization.
K-Nearest Neighbors (KNN)	Instance-based; non-linear boundaries.	Groups similar reservoirs (e.g., by porosity-permeability).
XGBoost	Gradient boosting with regularization; tabular data leader.	Proven in geoscience tasks.

2.11 Feature Engineering

To find the most important features that affect EOR success, we used a careful selection process. We removed repeated or less important features and kept only those that had a strong impact on the results [13]. This helped make the dataset easier to work with and improved the model's accuracy. First, we used a Heatmap as shown in Figure 5 to remove features that were highly related to each other because such features provide similar information, which can confuse the model or cause biased results.

2.12 Models Selection and Implementation

Selecting appropriate machine learning models is a critical step for the EOR screening task. Based on a comprehensive evaluation of model characteristics and dataset requirements, four algorithms were selected

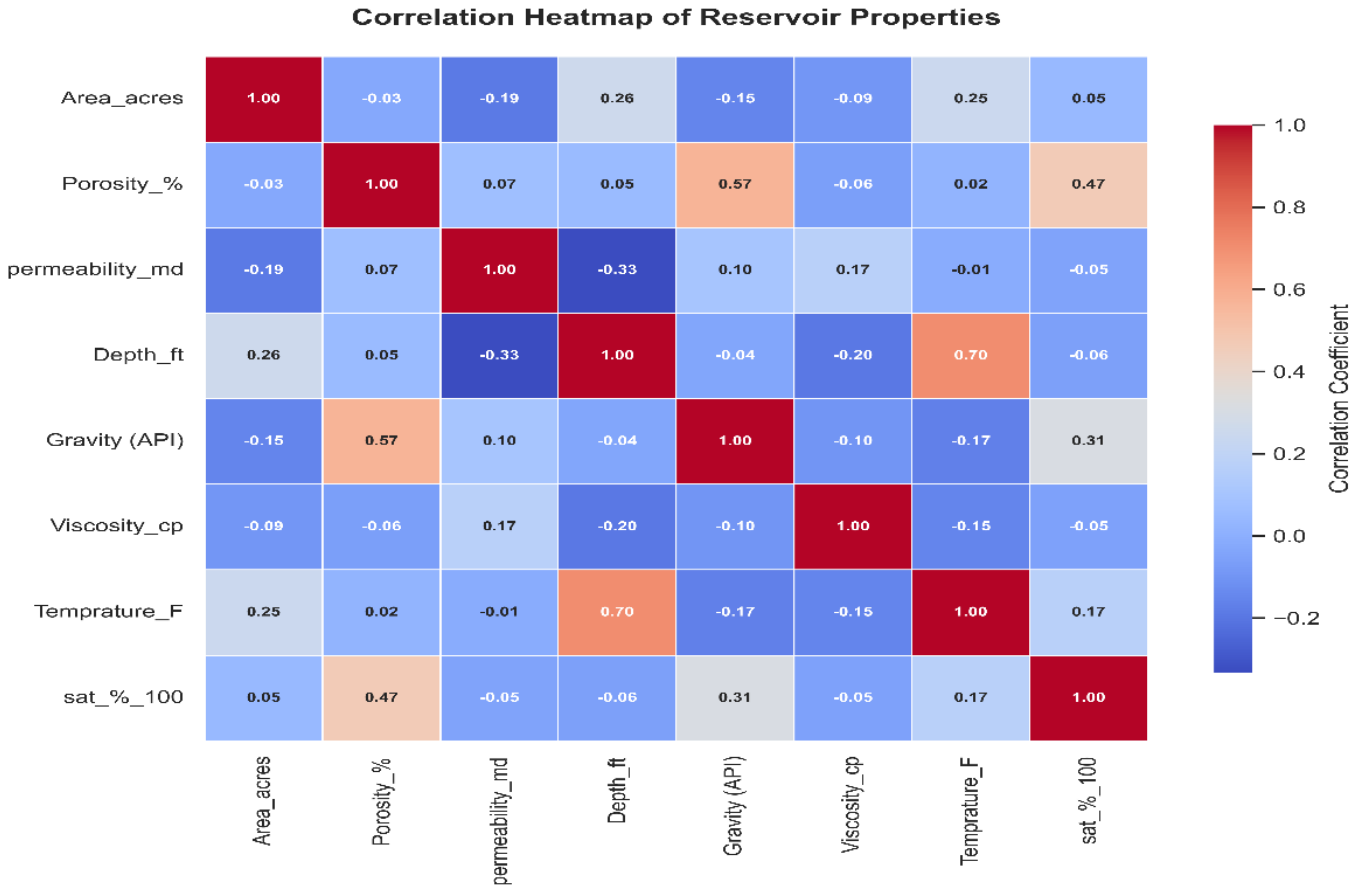


Figure 5. Heat map shows correlating features of Reservoir Properties.

for this study. Table 2 summarizes the rationale for this selection, outlining the key strengths of each model and their specific relevance to the EOR screening problem.

Picking the right machine learning models for EOR screening is crucial, we need solid predictions that petroleum engineers can actually understand and trust, without burning through compute resources on our lopsided dataset (10 EOR types, originally severe imbalance ratio of 123:1, now balanced to 760 samples after SMOTE).

Selection Criteria

Models were chosen based on:

- **Model selection prioritized practical performance** for EOR classification under real-world constraints like class imbalance and limited data [14].

Handling Multi-Class Imbalance

Targeted models that natively manage 10 EOR classes with extreme skew—CO_2 miscible at just 11 samples—without collapsing to majority-class bias.

Ensuring Interpretability

Required transparent decision logic that petroleum engineers can trace back to physical screening criteria like API gravity >22° and depth >2500 ft.

Dataset Compatibility

Focused on algorithms proven effective for modest tabular datasets (760 samples, 14 features) that need minimal tuning for reservoir engineering deployment.

Excluded Models

SVM: Multi-class complexity (45 classifiers); imbalance bias; low interpretability.

Neural Networks: Small dataset (760 samples); overfitting risk; tuning demands; black-box nature.

Logistic Regression: Linear assumptions unfit non-linear EOR relationships.

2.12.1 Evaluation Strategy

K-Fold Cross-Validation. For the evaluation method, we used 5-fold cross-validation (K=5). We split the dataset into 5 parts (folds). Each time, we trained the model on 4 folds (approximately 608 samples) and tested it on the remaining fold (approximately 152 samples). This process was repeated 5 times, with each fold serving as the test set exactly once. The K=5 value was selected to balance computational efficiency with reliable performance estimation, ensuring each test partition had adequate representation of all 10 EOR classes (15-20 samples per class per fold).

Rationale for Macro-Averaged Metrics: Given the class imbalance in our EOR dataset, we deliberately chose macro-averaged precision, recall, and F1-score (Table 4) rather than weighted or micro-averaging. Macro-averaging calculates metrics independently for each of the 10 EOR classes, then takes their unweighted mean. This ensures equal consideration of rare methods (e.g., CO_2 miscible with 60-70 samples) and common methods (e.g., polymer flooding with 110 samples), aligning with engineering practice where correctly identifying the optimal EOR method is equally critical regardless of reservoir type frequency in historical data. Weighted averaging would mask poor performance on minority classes, while micro-averaging would measure only overall accuracy without class-specific insight, neither appropriate for our imbalanced, multi-class EOR screening context [21].

2.13 EOR Screening Modeling and Model Selection

2.13.1 Decision Tree

We chose the Decision Tree method because it's easy to follow and understand [15]. It's like a flowchart that asks questions about the data one at a time and then groups the answers. The model's average accuracy was about 84% after the settings were changed. It got close to 80% each time we tested it with K-Fold cross-validation.

2.13.2 Random Forest

We picked Random Forest because it usually makes better predictions and helps us avoid making mistakes by overfitting [18]. This method builds a number of Decision Trees, each of which looks at a different random part of the data [19]. The accuracy was about 85% after fine-tuning, and the repeated K-Fold tests showed an 84% result.

2.13.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) was selected for modeling Enhanced Oil Recovery (EOR) because it works well with data that does not have any pattern. KNN decides the class of a point by looking at the classes of its closest neighbors and picking the most common one. After adjusting its settings for the best results, the KNN model achieved an average accuracy of 84%. This shows that it can effectively understand complex patterns in the data. It was tested using

Table 3. Optimized hyperparameters for machine learning models.

Model	Key Hyperparameters Tuned	Best Values Obtained
Decision Tree	Criterion, Max depth, Max features, Min samples leaf, Min samples split	Entropy, 20, sqrt, 1, 2
Random Forest	Criterion, Max depth, Max features, Min samples leaf, Min samples split, n estimators	Gini, None, sqrt, 1, 2, 100
KNN	Metric, neighbors, Weights	Euclidean, 3, Distance
Xgboost	Learning rate, Max depth, n estimators	0.1, 4, 300

K-Fold cross-validation, the model's average accuracy was 83%.

2.13.4 XGBoost

We used XGBoost for EOR screening because it can find important features and deal with complicated relationships in the data. It makes a strong overall model by putting together a lot of simple models, which are called weak learners. After using Grid Search cross-validation to fine-tune its settings, XGBoost got an average accuracy of 81%. Using K-Fold cross-validation, the model's accuracy was 80%.

2.13.5 Hyperparameter Optimization and Dimensionality Reduction

Hyperparameter Optimization

Hyperparameter optimization is essential for maximizing model performance [17, 23] in EOR screening. We employed Grid Search Cross-Validation to systematically identify optimal hyperparameters for each model. Grid Search exhaustively evaluates all possible combinations of specified parameter values, using 5-fold cross-validation to assess each combination's performance. The combination achieving the highest macro-average F1-score was selected as optimal.

The parameter grids tested for each model are presented below:

Decision Tree: criterion (entropy, gini), max_depth (10, 20, 30, None), max_features (sqrt, log2, None), min_samples_leaf (1, 2, 5), min_samples_split (2, 5, 10)

Random Forest: criterion (gini, entropy), max_depth (10, 20, 30, None), max_features (sqrt, log2, None), min_samples_leaf (1, 2, 5), min_samples_split (2, 5, 10), n_estimators (50, 100, 200)

KNN: n_neighbors (3, 5, 7, 9), metric (euclidean, manhattan, minkowski), weights (uniform, distance) XGBoost: learning_rate (0.01, 0.1, 0.3), max_depth (3, 4, 5, 6), n_estimators (100, 200, 300)

The optimal hyperparameters identified through Grid Search are reported in Table 3.

Dimensionality Reduction using PCA:

We applied Principal Component Analysis (PCA) for dimensionality reduction while retaining 95% of the dataset's information [16]. The correlation heatmap (Figure 4) revealed high correlations among several reservoir parameters, such as depth-temperature (due to geothermal gradient) and porosity-permeability relationships. This redundancy can cause overfitting and increases computational cost without improving predictions.

All features were standardized using StandardScaler (mean=0, variance=1) before PCA application, as PCA is scale sensitive. Without standardization, features with larger numerical ranges (depth in thousands of feet) would dominate over smaller-range features (oil saturation as fractions).

We applied PCA with the threshold set to retain 95% of the total variance, balancing information preservation with dimensionality reduction [20]. The PCA transformation was fitted exclusively on training data (608 samples), then applied to both training and test sets (152 samples) to prevent data leakage.

PCA reduced the feature space from seven original features while preserving 95% of information, decreasing training time by approximately 25-30% while maintaining model accuracy within 1-2% of the full feature set. However, PCA reduces interpretability since the transformed features are mathematical

combinations of original reservoir properties, making it harder to explain predictions in terms of specific parameters like depth or API gravity.

2.13.6 *Evaluation Metrics and Their Relevance to EOR Screening*

The imbalance in EOR datasets means that focusing only on accuracy doesn't give a clear picture of how well the models perform. That's why we concentrated on precision, recall, and F1-score as shown in Table 4, Specifically, we report macro-averaged metrics, where each EOR class receives equal weight regardless of sample size. This choice is critical because in reservoir engineering practice, accurate predictions for rare reservoir types (represented by minority classes) are as valuable as predictions for common types. Macro-averaging prevents our models from achieving high scores simply by favoring majority classes while performing poorly on underrepresented EOR methods. To better understand how effectively the models identify the less prevalent EOR approaches. We also took a thorough look at confusion matrices to detect where the model can misinterpret similar EOR strategies, like miscible vs immiscible gas injection. This extensive review technique guarantees we properly grasp the advantages given by our synthetic data production strategy.

Table 4. Comparative evaluation metrics (macro-averaged) for four ML models.

Model	Precision	Recall	F1-Score
Decision Tree	0.85	0.85	0.85
Random Forest	0.90	0.85	0.85
KNN	0.90	0.89	0.89
Xgboost	0.83	0.81	0.81

3 Results and Discussion

The models used to predict the best enhanced oil recovery (EOR) techniques for reservoirs were trained carefully. They were also fine-tuned to work well with the data. We looked closely at how making synthetic data affected their performance. This helped make sure the models gave accurate and useful predictions. The Decision Tree model performed well, benefiting from its simplicity in clearly identifying decision boundaries, as seen in Table 3. The Random Forest model showed a significant boost in accuracy after fine-tuning its hyperparameters, thanks to its ability to combine many Decision Trees and capture complex patterns in the data. The K-Nearest Neighbors (KNN) and Random Forest models were competitive, effectively recognizing data patterns through distance-based classification of closest neighbors and tree-based ensemble approaches, which is illustrated in Figure 6. Lastly, the XGBoost model, known for its gradient boosting technique, also delivered strong results, though its accuracy depended on the tuning and data specifics. Overall, these models demonstrated different strengths in predicting suitable EOR methods for reservoirs, with performance improvements visible after careful training and optimization [24].

3.1 Impact of Synthetic Data Generation

To showcase how well our Domain-Informed Synthetic Data Generation (DI-SDG) approach works, we pitted it head-to-head against the popular SMOTE technique. We applied both methods to balance the exact same starting dataset, ending up with 760 samples each time, and then trained and tested all four machine learning models using the very same setup and process.

3.2 Experimental Setup

To make sure we compare traditional SMOTE fairly against our new Domain-Informed Synthetic Data Generation (DI-SDG) method, we set up a controlled experiment using the exact same original dataset under identical conditions. After cleaning, the imbalanced dataset had 575 samples, and both techniques added 185 synthetic samples to create balanced sets of 760 samples each, so any differences in results came purely from the quality of the generated data, not the total size.

For SMOTE, we used the standard version with k=5 nearest neighbors, which picks random minority class examples and creates new ones by linearly blending them with their closest neighbors in feature space.

In contrast, our DI-SDG approach draws on EOR screening criteria from Table 1, generating samples through constrained random sampling that stays within valid ranges for each EOR type, followed by geological checks to confirm realistic reservoir conditions.

We tested both using 5-fold cross-validation with the same train-test splits every time, and tuned all four models,Decision Tree, Random Forest, KNN and XGBoost, with identical hyperparameter searches to keep everything apples-to-apples.

Table 5. Comparison of Model Performance Using SMOTE and Domain-Informed Synthetic Data Generation

Model	SMOTE Accuracy (Folds 1–5)	SMOTE Mean	DI-SDG Accuracy (Folds 1–5)	DI-SDG Mean	Improvement
Decision Tree	0.72, 0.70, 0.75, 0.78, 0.80	0.750	0.78, 0.74, 0.80, 0.84, 0.86	0.804	+5.4%
Random Forest	0.82, 0.77, 0.82, 0.84, 0.83	0.816	0.86, 0.80, 0.86, 0.88, 0.87	0.854	+3.8%
KNN	0.80, 0.80, 0.82, 0.84, 0.83	0.818	0.83, 0.83, 0.86, 0.88, 0.87	0.854	+3.6%
XGBoost	0.75, 0.75, 0.72, 0.84, 0.72	0.756	0.79, 0.79, 0.76, 0.88, 0.76	0.796	+4.0%
Average	–	0.785	–	0.827	+4.2%

3.3 Performance Comparison Result

Table 5 presents the k-fold cross-validation accuracy results for all four models trained on datasets augmented with SMOTE versus our DI-SDG method. Our domain-informed approach outperformed SMOTE for every single model tested, with improvements ranging from 3.6% to 5.4%. This consistency demonstrates that the benefit comes from the quality of synthetic data rather than model-specific effects.

3.4 Model Performance Analysis

KNN and Random Forest both achieved the highest accuracy of 0.854, significantly outperforming Decision Tree (0.804) and XGBoost (0.796). KNN's superior performance stems from its distance-based classification, which naturally aligns with EOR selection, reservoirs with similar characteristics (depth, API gravity, viscosity) require the same EOR technique. The balanced dataset was particularly beneficial for KNN, as the previous severe imbalance (123:1) meant minority class samples had very few same-class neighbors. After balancing 60-110 samples per class, KNN could reliably identify similar reservoirs within each category.

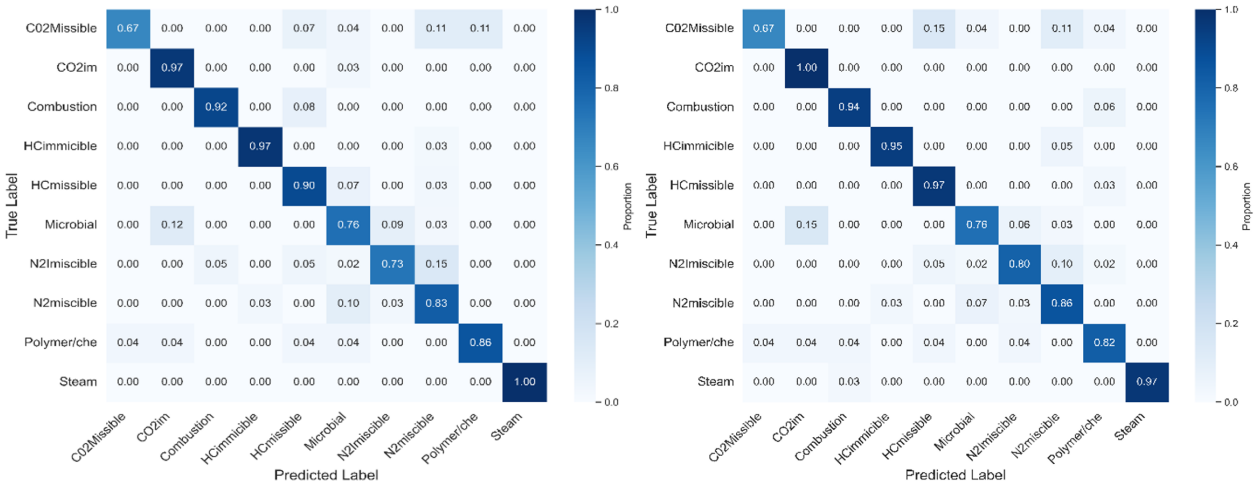
Random Forest excelled due to its ensemble approach and ability to capture complex feature interactions. By combining multiple decision trees, it reduces overfitting and automatically learns non-linear relationships critical to EOR selection, such as depth-temperature correlation and the inverse API-viscosity relationship. Additionally, it handles mixed data types (continuous and categorical variables) effectively without extensive preprocessing.

3.5 Confusion Matrix Analysis of Best Performing Models

To provide detailed insight into model classification behavior, we present confusion matrices for KNN and Random Forest (Figure 6(a,b)),the two models that achieved the highest accuracy (0.854). These matrices reveal how each model handles class boundaries and where misclassifications occur, offering insights beyond overall accuracy metrics. We focus on these two best-performing models as they represent the most reliable options for practical EOR screening deployment. Decision Tree (0.804) and XGBoost (0.796), while showing respectable performance, exhibited higher error rates and less consistent predictions across minority classes, making detailed confusion analysis of all four models redundant for practical insights.

Confusion Matrix Analysis: Co_2 Miscible vs. HCmiscible Gas Injection has been examined the confusion matrices for both KNN and Random Forest (Figure 6(a,b)) to understand model performance on boundary cases. Both models exhibit similar confusion patterns, with the main confusion occurring between both miscible gas injection methods.

CO_2 Miscible vs. HC miscible Confusion (11-15% error rate): This category has the highest misclassification rate among all methods. The most significant confusion occurs with HC miscible, where 11-15% of CO_2 miscible cases are incorrectly predicted as HC miscible. This confusion makes sense from a technical standpoint. Both CO_2 miscible and hydrocarbon miscible flooding work on similar principles, they both achieve miscibility between the injected fluid and reservoir oil when pressure is high enough [16]. The reservoir conditions needed for both methods are quite similar: sufficient depth to maintain high pressure (usually above 2500 ft), suitable oil characteristics (API gravity around 25-45°), and enough permeability for effective displacement. The main difference between them is just what fluid you're injecting, not the fundamental reservoir requirements. So when the models look at reservoir features like depth, permeability, temperature, and oil properties, it's not surprising they sometimes have trouble telling these two miscible flooding methods apart, especially when reservoir conditions could work for either approach.



(a) KNN (b) Random Forest

Figure 6. Confusion matrices for (a) KNN and (b) Random Forest.

Both KNN and Random Forest show similar confusion patterns here, with error rates of 7% (KNN) and 15% (Random Forest) for the CO_2Miscible and HC miscible misclassification. The fact that both models struggle in the same way suggests this confusion isn't just a quirk of one particular model, it reflects the real technical similarity and overlapping requirements between these two miscible EOR methods.

3.6 Limitations and Challenges

While this study demonstrates the effectiveness of domain-informed machine learning for EOR screening, we recognize several important limitations.

3.6.1 Dataset Limitations

Our dataset of 760 samples, though sufficient for proof-of-concept, remains modest. Even after augmentation, rare EOR methods have only 60-110 samples each, which may not fully capture reservoir variability. The dataset also shows geographical bias, projects from the Middle East, North Sea, or South America are underrepresented compared to North American cases, potentially limiting model performance in different geological settings.

We use only seven reservoir parameters, but EOR success depends on many additional factors we couldn't include: oil composition (asphaltene content, sulfur levels), formation characteristics (heterogeneity, fractures, geological structure), and operational factors (well spacing, infrastructure). These missing features likely explain some confusion between similar EOR methods. Additionally, our data comes from published projects, which favors successful cases since failures are less reported, creating success bias in our training data.

3.6.2 Methodological Limitations

Our synthetic data generation, while better than SMOTE, constrains samples to Table 1 parameter ranges, which may limit handling of truly novel reservoir conditions. We also make simplifying assumptions about parameter independence that don't fully capture complex reservoir interdependence.

Random Forest and KNN achieve 85.4% accuracy but lack transparency it's difficult to explain exactly why specific predictions were made. This could be problematic in industry settings where engineers must justify recommendations to management. There's an inherent trade-off: simpler models are easier to understand but less accurate.

Our k-fold cross-validation tests performance on historical data but lacks prospective validation testing predictions on new reservoirs before EOR selection. Real-world performance on future screening may differ from our 85.4% accuracy. Our metrics also treat all mistakes equally, when recommending a completely wrong method is far worse than confusing similar ones.

3.6.3 Practical Implementation Challenges

A major difference is the absence of economic analysis. Our models focus on technical feasibility, but real EOR decisions depend heavily on economics: oil prices, injection costs, infrastructure requirements, and profitability. Technically perfect methods might be economically unfeasible.

Another challenge is temporal validity. EOR technology evolves continuously, but our models don't automatically adapt to improvements unless retrained. Similarly, changing oil prices, environmental

regulations, and carbon pricing alter economic feasibility in ways our static models can't capture.

We also face uncertainty about performance on unconventional reservoirs (tight oil, shale) or emerging techniques not in our training data. Our conservative approach to respecting established criteria might limit recognizing when methods could work in novel situations.

3.6.4 Future Research Challenges

Looking ahead, we face several challenges. Acquiring comprehensive data is difficult, many valuable projects remain proprietary. Building industry partnerships for broader datasets while respecting confidentiality will be crucial. We also need to integrate information across scales, from pore-level fluid behavior to field-scale production.

Another gap is uncertainty quantification our predictions lack rigorous confidence intervals. When we report 78% probability, we can't distinguish confident predictions from uncertain ones [22]. Finally, developing models that learn continuously from new field data while maintaining stability presents technical challenges.

3.7 Opportunities

Despite these limitations, this research opens significant opportunities for EOR practice and petroleum engineering.

3.7.1 Enhanced Decision-Making and Industry Applications

The 85.4% accuracy demonstrates machine learning can genuinely support EOR screening. This enables comprehensive decision-support systems combining technical screening with economic models, risk assessment, and project planning transforming EOR screening from expert-dependent judgment into transparent, data-driven methodology while preserving engineering insight when needed.

Our domain-informed approach for handling imbalanced data applies beyond EOR screening. Many petroleum problems involve imbalanced datasets rare reservoir types, uncommon drilling problems, infrequent equipment failures. This methodology could extend to well log synthesis for rare rock types, production curves for new reservoirs, or drilling optimization with scarce data.

3.7.2 Advancing Machine Learning and Integration

While Random Forest and KNN perform well, opportunities exist for more sophisticated approaches. Deep learning might capture complex patterns we're missing. Graph neural networks could represent reservoir connectivity explicitly. Transfer learning could leverage data-rich situations to improve data-poor ones. Explainable AI techniques like SHAP could make models transparent without sacrificing accuracy, achieving both better predictions and better understanding.

We could also connect screening more tightly with reservoir simulation. Machine learning identifies promising candidates for detailed simulation, then simulation results improve screening models, creating continuous improvement cycles. Multi-objective optimization could simultaneously consider technical performance, economics, environmental impact, and operational risk.

3.7.3 Expanding Scope and Continuous Learning

Our methodology provides foundation for evaluating emerging EOR techniques (low-salinity waterflooding, nanoparticle recovery) with limited field history and adapting to unconventional reservoirs where traditional criteria don't apply. This is particularly relevant as industry focuses on unconventional assets and carbon-neutral approaches.

Instead of one global model, regional variants trained on basin-specific data (Middle East carbonates, North Sea sandstones, Gulf Coast heavy oil) could achieve higher accuracy by learning local patterns and practices. Federated learning could enable collaborative improvement while maintaining data privacy.

The web application we developed (Section 4) enables learning from actual usage. As engineers report outcomes selected methods and project performance feedback could systematically improve models. This creates systems that get better with use, addressing temporal validity by staying current with evolving technology.

3.7.4 Broader Impact

The core idea is to use domain knowledge to guide machine learning with limited data extends beyond petroleum engineering. Geothermal energy, groundwater remediation, CO_2 sequestration, and mining face similar challenges. The framework could transfer to these fields, creating broader impact.

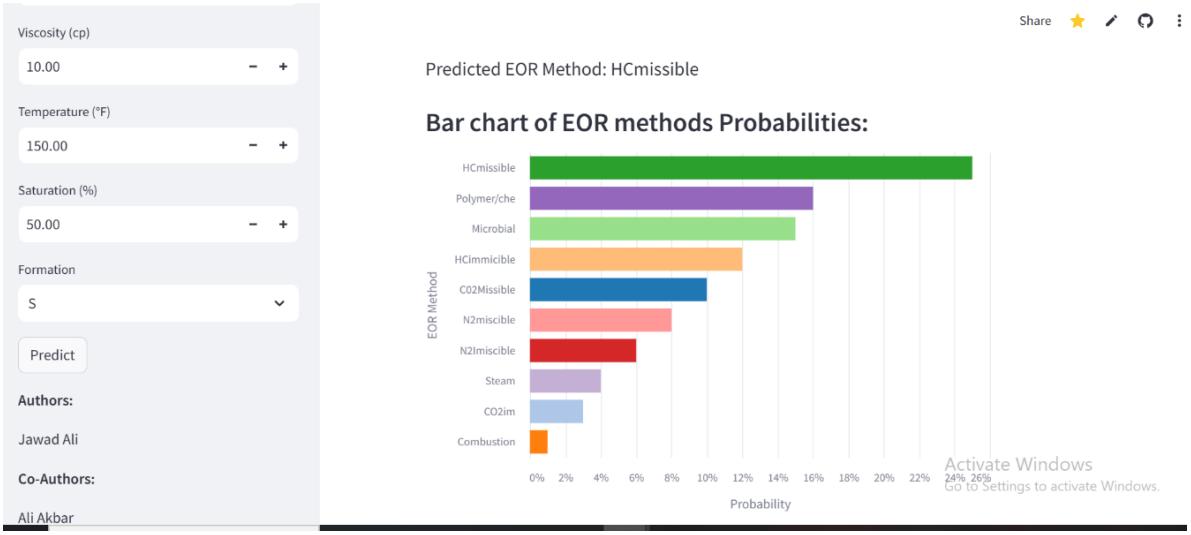


Figure 7. Streamlit application interface showing input sidebar (left) and output visualization with bar chart and probability table (right).

Finally, our tool offers educational opportunities. Students can interactively explore how reservoir parameters influence EOR selection, developing intuition for screening criteria while learning to integrate traditional petroleum engineering with modern data science approaches.

3.8 Future Work

3.8.1 Practical Tool Implementation and Future Development

As part of our third contribution, we have developed a proof-of-concept web application using Streamlit that demonstrates the practical deployment of our trained machine learning models for EOR screening.

3.8.2 Current Implementation: Streamlit Web Application

We developed a functional web-based application accessible via browser without installation. Users input eight reservoir parameters (Depth, API Gravity, Viscosity, Temperature, Oil Saturation, Permeability, and Formation Type, Area in Acres) through an interactive sidebar with sliders and dropdown menus. Upon clicking the "Predict" button, the application displays results in two formats: a probability table listing all ten EOR methods with their prediction probabilities in descending order, and a color-coded horizontal bar chart providing visual comparison as demonstrated in Figure 7. The application clearly indicates the predicted EOR method with the highest probability.

Predicted EOR Method: HC miscible (25%)

Ranked Probabilities: HC miscible (25%), Polymer/Chemical (16%), Microbial (15%), HC immiscible (12%), CO_2Miscible (10%) still remaining methods.

3.8.3 Current Scope and Limitations

This research prototype demonstrates feasibility but has limitations: limited training dataset (760 samples), no economic analysis, single reservoir screening only, no detailed rationale or screening criteria validation, and no field validation yet. It serves as proof-of-concept rather than a production-ready system.

3.8.4 Future Development Vision

Near-term: Add screening criteria compliance checking, detailed rationale for recommendations, batch processing for multiple reservoirs, and downloadable reports.

Mid-term: Expand training dataset to >2000 samples, develop regional model variants, integrate economic screening module with cost analysis.

Long-term: Develop API for integration with petroleum software (Petrel, Eclipse, CMG), create mobile application, implement field validation feedback loop for continuous model improvement.

Future deployment options include cloud-based web service, on-premises installation for data security, or embedded plugin for existing reservoir engineering software.

4 Conclusion

This study represents a significant advancement in enhanced oil recovery (EOR) screening by leveraging

state-of-the-art machine learning techniques. The research yields several key findings and contributions. The developed and optimized models—Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost—demonstrated distinct strengths and limitations, yet collectively enhanced the interpretability and reliability of EOR method selection for reservoirs. Notably, the strategic application of domain-informed synthetic data generation effectively mitigated severe class imbalance, resulting in a more balanced and representative dataset. Overall, the findings highlight the practical potential of machine learning to improve the screening and selection of EOR techniques. The integration of dimensionality reduction, hyperparameter optimization, and constrained synthetic data augmentation has led to the construction of more robust and dependable predictive models. In conclusion, this work lays a strong foundation for more intelligent, data-driven decision-making in enhanced oil recovery. As the oil and gas industry continues to evolve, such machine learning-driven approaches are poised to play an increasingly vital role in optimizing planning, operational efficiency, and resource recovery.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Aladasani, A., & Bai, B. (2010, June). Recent developments and updated screening criteria of enhanced oil recovery techniques. In *SPE International Oil and Gas Conference and Exhibition in China* (pp. SPE-130726). Spe. [CrossRef]

[2] Cheraghi, Y., Kord, S., & Mashayekhizadeh, V. (2021). Application of machine learning techniques for selecting the most suitable enhanced oil recovery method; challenges and opportunities. *Journal of Petroleum Science and Engineering, 205*, 108761. [CrossRef]

[3] Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2), 34-38.

[4] Alvarado, V., Ranson, A., Hernandez, K., Manrique, E., Matheus, J., Liscano, T., & Prosperi, N. (2002, October). Selection of EOR/IOR opportunities based on machine learning. In *SPE Europec featured at EAGE Conference and Exhibition?* (pp. SPE-78332). SPE. [CrossRef]

[5] Wong, T. T., & Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 32*(8), 1586-1594. [CrossRef]

[6] Al Adasani, A., & Bai, B. (2011). Analysis of EOR projects and updated screening criteria. *Journal of Petroleum Science and Engineering, 79*(1-2), 10-24. [CrossRef]

[7] Oil & Gas Journal. (1998, April 20). 1998 worldwide EOR survey [Industry survey]. Retrieved from https://www.ogj.com/home/article/17226236/1998-worldwide-eor-survey

[8] Taber, J. J., Martin, F. D., & Seright, R. S. (1997). EOR screening criteria revisited Part 1: Introduction to screening criteria and enhanced recovery field projects. *SPE Reservoir Engineering, 12*(3), 189-198. [CrossRef]

[9] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020, April). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)* (pp. 243-248). IEEE. [CrossRef]

[10] Provost, F. (2000, July). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (Vol. 68, No. 2000, pp. 1-3). AAAI Press.

[11] Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC. [CrossRef]

[12] May, R. J., Maier, H. R., & Dandy, G. C. (2010). Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks, 23*(2), 283-294. [CrossRef]

[13] Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems, 66*(3), 1575-1637. [CrossRef]

[14] Hartono, A. D., Hakiki, F., Syihab, Z., Ambia, F., Yasutra, A., Sutopo, S., ... & Apriandi, R. (2017, October). Revisiting EOR projects in Indonesia through integrated study: EOR screening, predictive model, and optimisation. In *SPE Asia Pacific Oil and Gas Conference and Exhibition* (p. D012S036R029). SPE. [CrossRef]

[15] Khazali, N., Sharifi, M., & Ahmadi, M. A. (2019). Application of fuzzy decision tree in EOR screening assessment. *Journal of Petroleum Science and Engineering, 177*, 167-180. [CrossRef]

[16] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access, 8*, 54776-54788. [CrossRef]

[17] Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges* (pp. 3-33). Cham: Springer International Publishing. [CrossRef]

[18] Frederick, L. (2005). Implementation of Breiman's Random Forest Machine Learning Algorithm. *ECE591Q Machine Learning Journal Paper*, 1-13.

[19] Parada, C. H., & Ertekin, T. (2012, March). A new screening tool for improved oil recovery methods using artificial neural networks. In *SPE western regional meeting* (pp. SPE-153321). SPE. [CrossRef]

[20] Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

[21] Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big data, 7*(1), 70. [CrossRef]

[22] Tarrahi, M., Afra, S., & Surovets, I. (2015, October). A novel automated and probabilistic EOR screening method to integrate theoretical screening criteria and real field EOR practices using machine learning algorithms. In *SPE Russian Petroleum Technology Conference* (pp. SPE-176725). SPE. [CrossRef]

[23] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295-316. [CrossRef]

[24] Zhang, N., Wei, M., Fan, J., Aldhaheri, M., Zhang, Y., & Bai, B. (2019). Development of a hybrid scoring system for EOR screening by combining conventional screening guidelines and random forest algorithm. *Fuel, 256*, 115915. [CrossRef]

Jawad Ali received his bachelor's degree in petroleum engineering from Mehran University of Engineering and Technology (MUET), Pakistan. His research interests lie at the intersection of Artificial Intelligence and Geoscience, with a particular focus on applying data-driven methods to subsurface characterization and carbon capture and storage (CCS) technologies. He has worked as a Research Assistant in the Department of Petroleum Engineering at MUET, where he contributed to projects integrating machine learning with geological and petrophysical data to enhance understanding of reservoir behavior and energy transition solutions. (Email: jawadaliv28@gmail.com)

Ubedullah Ansari is an Associate Professor at the Institute of Petroleum & Natural Gas Engineering, Mehran University of Engineering & Technology (MUET), Jamshoro, Pakistan. He holds a PhD in Oil and Gas Well Engineering and has more than fourteen years of academic and professional experience in petroleum engineering, reservoir simulation, and energy optimization. His research interests encompass geothermal energy extraction from subsurface reservoirs, shale gas characterization, enhanced oil recovery, CO_2 sequestration, and AI-driven digital twin modeling for reservoir management. Beyond academia, Dr. Ansari is the Founder of OG-Fleet, an AI-based platform for petroleum professionals integrating generative AI and virtual reservoir twins. He continues to mentor graduate researchers and industry trainees while contributing to Pakistan's transition toward sustainable and intelligent energy systems. (Email: Ubedullah.ansari@faculty.muet.edu.pk)

Fateh Ali, from Pakistan, holds a bachelor's degree in Petroleum Engineering from Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan. He is currently pursuing a master's degree at Chengdu University of Technology, Sichuan, China. During his time at MUET, he served as a Research Assistant at the Institute of Petroleum and Natural Gas Engineering (IPNGE). His research interests and experience encompass reservoir engineering, enhanced oil and gas recovery, and reservoir simulation. He is also conducting experimental research on underground hydrogen storage. (Email: fatehhingorjo@gmail.com)

Tariq Javed holds a bachelor's degree in petroleum engineering from Mehran University of Engineering and Technology (MUET), Pakistan. His academic and professional interests focus on reservoir engineering, enhanced oil recovery, and sustainable energy solutions. He has completed multiple field internships with leading oil and gas companies, including Oil and Gas Development Company Limited (OGDCL) and Pakistan Petroleum Limited (PPL), where he gained hands-on experience in reservoir characterization, well testing, and production optimization. He is passionate about applying innovative technologies and data-driven methods to improve reservoir performance and contribute to the future of energy transition. (Email: Engr.tariq1917@gmail.com)

Imran Ahmed Hullio received his degree in Petroleum and Natural Gas Engineering from Mehran University of Engineering and Technology, Jamshoro, Pakistan. His areas of specialization include hydraulic fracturing, reservoir engineering, and production engineering. His primary research interests focus on challenges associated with CO_2 injection and hydraulic fracturing. (Email: imran.hullio@faculty.muet.edu.pk)