Predicting performance of in-situ microbial enhanced oil recovery process and screening of suitable microbe-nutrient combination from limited experimental data using physics informed machine learning approach

Sree Pavan[a], K. Arvind[b], B. Nikhil[b], P. Sivasankar[a,*]

[a] Geo-Energy Modelling & Simulation Lab, Department of Petroleum Engineering, Indian Institute of Petroleum & Energy, Visakhapatnam - 530003, India.

[b] Department of Mechanical, Chemical and Electronics Engineering, OsloMet University, Oslo, Norway

[*]Corresponding author

Contact Details of Corresponding Author (P. Sivasankar)

Email ID: sivasankar.petro@iipe.ac.in

Ph: 91 9600043460

[Original Research Article]

Submitted to

Bioresource Technology

January 2022.

DOI: https://doi.org/10.1016/j.biortech.2022.127023

Abstract

To screen/identify suitable microbe, nutrient and reservoir for successful field implementation of in-situ MEOR technique, it is important to predict the oil recovery and quantify the relative importance of influencing parameters from limited experimental data. For this purpose, Physics-Informed Machine Learning (PIML) approach is adopted in this study, which is developed by integrating the physics-based and Machine Learning (ML) models. It is found that biosurfactant yield *w.r.t* nutrient (Y_{PS}), flow velocity and initial oil saturation (S_{oi}) are correspondingly the most influential microbial kinetic, operational and reservoir parameters. Higher oil recovery is achieved by selecting a microbe-nutrient-reservoir pair having higher Y_{PS} and S_{oi} values but with lower Y_{XS} (microbial yield w.r.t nutrient) value. Among 12 ML models analysed, Neural network model had predicted the oil recovery relatively accurate ($R^2 \sim 0.98$). Overall, this PIML approach helps to devise strategies for maximizing oil recovery at initial laboratory stage itself with limited experimental data.

Keywords: Microbial Enhanced Oil Recovery; Machine Learning; Biosurfactants; Modelling; Kinetics

1. Introduction

To meet the increase in global energy demand and to sustain crude oil production from depleting oil reservoirs, more than half of the crude oil that is left after primary and secondary recovery techniques must be recovered by suitable Enhanced Oil Recovery (EOR) techniques (Joshi et al., 2016). In relative to existing chemical EOR methods, In-situ Microbial Enhanced Oil Recovery (MEOR) method is an economical and environmental friendlier EOR technique (Joshi et al., 2016; Varjani and Upasani, 2016; Shibulal et al., 2018; Jeong et al., 2022). In in-situ MEOR process, exogeneous (or) indigenous microbes are injected into the reservoir, which subsequently undergoes metabolic activity within the reservoir by utilizing nutrients and producing bioproducts, which consequently helps to recover the crude oil from the reservoirs (Joshi et al., 2016; Varjani and Upasani, 2017; Shibulal et al., 2018; Markande et al., 2021). Though in-situ MEOR technique inherits several advantages, it is not widely implemented in the field across the globe as other chemical EOR techniques due to the existence of following challenges (Nikolova and Gutierrez, 2020): (a) complexity in predicting the oil recovery performance of in-situ MEOR technique; and (b) lack in quantifying the relative importance of each influencing parameter on final oil recovery. Resolving these challenges at initial lab investigation stage itself will correspondingly: help to decide whether to implement in-situ MEOR technique in the given reservoir or not and to identify/screen the suitable microbe-nutrient-reservoir combination for attaining better oil recovery; and assist in development of strategies for optimizing the oil recovery.

To evaluate the oil recovery performance of in-situ MEOR process, earlier, several core flooding experimental studies (Joshi et al., 2016; Varjani and Upasani, 2016; Shibulal et al., 2018) and physics based computational modelling studies (Nielsen et al., 2016; Sivasankar and Kumar, 2016, 2017, 2019; Jeong et al., 2021, 2022) were performed. However, performing core flooding experimental studies to identify/screen a suitable microbe-nutrient-reservoir combination from several available combinations makes experimental approach an expensive and time-consuming exercise. Though physics-based models can provide better prediction of oil recovery with physically consistent results, but it is computationally intensive to perform uncertainty quantification and optimization studies as it requires to solve the non-linear equations for several simulation runs (Thanh et al., 2020; Karniadakis et al., 2021). Moreover, it is also unfeasible to quantify the relative importance of each influencing parameter on oil recovery by both experimental and physics-based modelling approach as it requires several experiments. Recently, Machine Learning (ML) models/algorithms are increasingly used for different bioprocess applications to predict and optimize its performance (Cruz et al., 2012; Tang et al., 2021; Zhang et al., 2021; Wang et al., 2022). With the availability of large input and output datasets, ML models can quickly predict the outcome of complex problems and quantify the relative importance of each input parameters, which is otherwise difficult by using only physics-based models (Thanh et al., 2021; Tang et al., 2021). However, with the limited availability of data from experimental and field studies, it will not be feasible to apply ML models/algorithms alone as it may predict physically inconsistent results with lesser accuracy. Hence the requirement to have a quick and physically consistent results from limited observed/experimental data with better accuracy is achieved by integrating both the physics informed model and data driven ML model into a single hybrid model called Physics Informed Machine Learning (PIML) model (Thanh et al., 2020; Karniadakis et al., 2021). In recent times, PIML modelling approach is gaining popularity because of its ability to accommodate the merits of both physics-based model and ML model in a single model, while mitigating their respective drawbacks. Recently, PIML modelling approach have been successfully used for different applications (Thanh et al., 2020; Karniadakis et al., 2021; Liu et al., 2021). However, the use of PIML approach for in-situ MEOR application have not been explored yet at least to the authors knowledge.

SOMETHING ABOUT PIML WRITE ABOUT EXPLAINABLITY, INTERPRETABLE AND PHYSICALLY CONSISTENT…

Hence the novelty of the present work is in introducing the PIML approach for in-situ MEOR application to predict its oil recovery performance and to quantify the relative importance of parameters influencing the oil recovery using limited experimental data. In particular, the objectives of the present work are: (a) to develop a framework to integrate the physics based model and ML model into a single PIML model for generating a large, relevant and physically consistent data sets from limited experimental data; (b) to quantify the relative importance of each parameter on influencing the final oil recovery using PIML approach, and subsequently to identify the critical kinetic and operational parameters influencing the oil recovery; and (c) to identify the suitable ML model among 12 different ML models that shall be used directly in PIML approach for predicting the oil recovery performance.

The present PIML approach study will help the end-users: to quickly select a favourable microbial-nutrient-reservoir combination from several other available options; to decide whether to implement in-situ MEOR technique in a particular reservoir or not; and to devise operational strategies for maximizing the oil recovery.

2. Materials and Methods

In the present study, PIML approach is developed by combining the physics-based model and ML model into a single model. Initially, laboratory experiments are performed to determine the microbial kinetic and reservoir properties data. Based on these limited experimental data, physics-based models for microbial kinetic and oil recovery processes are developed. From these physics-based models, large, physically relevant input and output data sets are generated. Using these large datasets, the ML models are then trained and tested to quantify the relative importance of each parameter and to predict the oil recovery quickly. The methodology for developing this PIML approach is presented in detail in this section and briefed in Figure 1.

[Figure 1]

2.1 Classification and collection of input parametric data

In the present study, 13 input parameters are considered. The corresponding values of these input parameters constitutes the input parametric data. In the present study, the input parametric data are classified as: (i) microbial kinetic parametric data, (ii) operational parametric data, and (iii) reservoir parametric data, based on the corresponding properties of microbes, nutrients, operational and reservoir conditions.

2.1.1 Collection of input microbial kinetic parametric data from experimental studies

In the present study, the microbial kinetic parameters that are considered as input are maximum microbial growth rate [$U_{max}, (h^{-1})$], yield of microbes *w.r.t* sucrose (Y_{XS}), yield of biosurfactants *w.r.t* sucrose (Y_{PS}) and Monod half saturation coefficient ($K_{XS}, (gl^{-1})$]. The corresponding values of these microbial kinetic parameters are considered as input microbial kinetic parametric data. In the present study, the input data for all these microbial kinetic parameters are sourced from the experimental studies of Sivasankar et. al., 2016, in which, *Pseudomonas putida* MTCC 2467 was used as microbe, while sucrose and ammonium sulphate were used as carbon and nitrogen source nutrient, respectively. In that study, at pH 8 condition, experiments on microbial growth, nutrient utilization and biosurfactant production were carried out to determine the values of microbial kinetic parameters for predicting the oil recovery.

2.1.2 Collection of input operational and reservoir parametric data

In the present study, the operational parameters that are considered as input are mean flow velocity of injection fluid within reservoir [$u, (mh^{-1})$], viscosity of injection fluid [$\mu_w, (Nhm^{-2})$], initial/injection concentration of microbes [$X_i, (gl^{-1})$], initial/injection concentration of sucrose [$S_i, (gl^{-1}$]], initial/injection concentration of ammonium sulphate [$A_i, (gl^{-1}$],] and resident time [$T_r, (h)$]. These input operational parameters are controlled by the operators/scientists in the field/laboratory during the implementation of in-situ MEOR technique. Finally, the reservoir fluid-rock parameters that are considered as input parameters in the present study are initial residual oil saturation [S_{ori}, (*fraction)*], irreducible water saturation [S_{wir}, (*fraction)*] and initial or maximum oil-water Interfacial Tension (IFT) at the start of EOR [$\sigma_{max}, (mNm^{-1})$]. In the present study, the input data for all these operational and reservoir rock-fluid parameters are sourced from Sivasankar et al., 2016. Table 1 presents the sourced data or reference value of all these input parameters. It is to be noted that for each input parameter, only one reference value is available either from experiments or other sources, which will be insufficient for applying the ML algorithms.

[Table 1]

2.2 Generation of large input and output datasets from physics-based model

In the present study, percent of oil recovery is the only parameter considered as output parameter. This output oil recovery parameter is influenced by all the input parameters (Sivasankar et al., 2016) that are mentioned in Tab. 1. In order to apply

Machine Learning (ML) algorithms for predicting the output oil recovery, large data sets of input and output parameters are required to train and test the ML algorithms. However, the availability of input and output parametric data from laboratory experiments and other sources are limited (as presented in Tab. 1), which is inadequate to implement the ML algorithms. Hence in the present study, large datasets of input and output parameters are generated synthetically (data augmentation) for training and testing the ML algorithms. Data augmentation is a mathematical method to synthesise more data from the known (experimental) data when there is data insufficiency. The methodology adopted in the present study for generation of input and output data is similar to the method earlier adopted by Thanh et al., 2020, and it is outlined in sec. 2.2.1. and sec. 2.2.2.

2.2.1 Generation of large input datasets from sourced reference values

The reference value of input microbial kinetic parameters that are presented in Tab. 1 are specific only to a particular temperature, pH, salinity, and pressure conditions at which experiments were conducted. However, in actual reservoir fields, the reference value of input parameters mentioned in Tab. 1 varies significantly due to the existence of heterogeneity, resulting in uncertainty (Ansah et al., 2020; Thanh et al., 2020). Hence accounting for this uncertainty, and to make the present model to be applicable for wider variations in input parametric data during its field implementation, a 50% Standard Deviation (SD) is considered to all the input parameter values (Thanh et al., 2020). The resultant value range for each of these input parameters after considering the SD is presented in Tab. 1. Subsequently, large datasets of about 10000 values (i.e., data) for each of the input parameter is generated between their corresponding value range by dividing it in equal intervals following the uniform distribution.

2.2.2 Generation of large output data sets using physics-based model

The output data on oil recovery is dependent on all the input parametric data. Hence, to generate the large data sets of output parameter (oil recovery, %) for training and testing of ML algorithms, the physics-based model (Eqs. 1- 8) (Sivasankar and Suresh Kumar, 2016) which is dependent on all input parameters is simulated several times using the generated input datasets. In the present physics-based model, the microbial kinetic model (Eqs. 1 – 4) simulates: growth kinetics of microbes (Eq. 1); nutrient utilization kinetics (Eq. 2); biosurfactant production kinetics (Eq. 3); and Monod's kinetics (Eq. 4). While the oil recovery model (Eqs. 5 – 8) simulates: IFT reduction by produced biosurfactants (Eq. 5); increase in Capillary Number due to IFT reduction (Eq. 6); decrease in oil saturation due to increase in Capillary Number (Eq. 7); and the final percent of oil recovery (Eq. 8), which is the output and target data. Based on this obtained oil recovery data, performance evaluation of MEOR technique and screening of suitable microbe-nutrient-reservoir combination are carried out.

$$dX/dt = \mu_x . X \quad (1)$$

$$dS/dt = -\mu_x . X/Y_{XS}; \qquad dA/dt = -\ \mu_x . X/Y_{XA} \quad (2)$$

$$dP/dt = (Y_{PS}/Y_{XS}) . \mu_x . X \quad (3)$$

$$\mu_x = U_{max} . \{(S/K_{XS} + S) + (A/K_{XA} + A)\} \quad (4)$$

$$\log(\sigma^*) = \log(\sigma_{min}) + \log(\sigma_{max}/\sigma_{min}) \ . \{(P - P_{max})/(P_{max} - P_{min})\} \quad (5)$$

$$N_{ca}^* = \ u_w \mu_w / \sigma^* \quad (6)$$

$$S_o = \left[\frac{-\tanh(v_1(N_{ca}^*) - v_3) + 1 + v_2}{-\tanh(v_1(N_{ca}^0) - v_3) + 1 + v_2}\right] S_o' \qquad S_w = 1 - S_o \quad (7)$$

$$Oil\ recovery, \% = \{(S_w - S_{wir})/(1 - S_{wir})\} \times 100 \quad (8)$$

In Eqs. (1 - 8), the terms: *X*, *S*, *A* and *P* represents the concentration of microbes, sucrose ammonium sulphate and produced biosurfactant, respectively in gl^{-1}; μ_x represents the microbial growth rate in h^{-1}; K_{XA} represents the half-saturation constant of ammonium sulphate in gl^{-1}; Y_{XA} represents the yield of microbes *w.r.t* ammonium and sulphate, r; Y_{PS} represents the yield of biosurfactant *w.r.t* sucrose; N_{ca} represents the updated IFT (mNm^{-1}) and Capillary Number, respectively; P_{min} and P_{max} represents the minimum and maximum biosurfactant concentration, respectively in gl^{-1}; σ^* and σ_{min} represents the updated IFT and minimum IFT, respectively in (mNm^{-1}); S_o and S_w represents the saturation of oil and water, respectively in fraction; and ν_1, ν_2, ν_3 are the constants.

By performing one simulation job of physics-based model from Eqs. 1 - 8, one output data on oil recovery is generated. Hence, in the present study, to generate a large database of output data, ten thousand (10000) simulation jobs were performed which resulted in generation of 10000 output data on % oil recovery. While, in each simulation job, the input value (data) of different input parameters that are required are sampled randomly from the generated input database using Latin Hyper-Cube Sampling (LHS) technique (Thanh et al., 2020). In some simulation jobs, the set of input data have not generated a valid positive output data (i.e., % of oil recovery), and such data are excluded from the training and testing of ML algorithms. Figure 2 shows the frequency distribution of all the input and output data values that were considered in the present study for training and testing of different ML algorithms.

[Figure 2]

2.3 Machine Learning Models

Subsequent to the generation of large sets of input and output data, the interaction strength (or) sensitivity of all 13 input parameters on the output oil recovery is quantified by using Pearson Correlation Coefficient (PCC) and Spearman Rank Corelation Coefficient (SRCC) values. PCC value measures the linear relationship between two variables and SRCC value measures the monotonic relationship between two parameters. Both PCC and SRCC values range from -1 to 1. Positive correlation value between two parameters indicates that parameters are directly proportional, and *vice versa*. Magnitude of the correlation indicates the strength of the relationship between the two parameters. Higher the magnitude of correlation coefficient value, higher is the association strength between the two parameters. Determination of PCC and SRCC values helps to quantify the influence of different input parameters on output oil recovery, which shall be used to screen the suitable microbes, nutrients and reservoirs at the laboratory experimental stage for MEOR field implementation.

2.4 Prediction of relative importance score to quantify the significance of input microbial kinetic, operational and reservoir parameters on output oil recovery

In the present work, feature importance study is carried out to quantify the relative importance of each input parameter on the output oil recovery using Random Forest Classifier ML algorithm (Keprate and Ratnayake, 2019) in the present PIML framework. This ML algorithm has been trained and tested using the input and output parameter datasets that are generated from physics-based model (as described in sec 2.3). This feature importance study computes the Relative Importance (RI) score for each input parameters in fraction, where its summation will be 1. Hence, RI score of an input parameter quantifies the significance (or) importance of that input parameter on influencing the output oil recovery in relative to other input parameters. Determination of this RI score for all input parameters will helps to identify the critical input parameters influencing the output oil recovery, which subsequently guide the future operation. In the present work, the results from the feature importance study would helps: (a) to identify the input parameters that are most and least important for predicting the oil recovery, which subsequently helps to identify the input parameters which exhibits higher and lower influence on the output oil recovery; (b) to identify the critical input parameters that shall be optimized for improving the efficiency of oil recovery; and (c) to determine the weightage functions of all input parameters, which shall be used to screen the suitability of MEOR technique among other EOR techniques and to identify the right combination of microbe-nutrient pair for attaining better oil recovery during its field implementation.

2.5 Prediction and evaluation of different machine learning algorithms for MEOR applications from lab data

In the present study, Machine Learning (ML) model which is integrated within the PIML approach is used to predict the output oil recovery. CRISP-DM methodology was used for performing data mining and predicting the output oil recovery from input parameters (Keprate and Ratnayake, 2019). The large data sets of input parameter data and output data that are required for training and testing the ML model are sourced from physics-based model which is embedded within the PIML approach (the procedure for data generation using physics-based model is presented in sec. 2.2). As there are different ML models available to do the prediction, it is necessary to identify the most accurate and suitable ML model that can be used in the PIML approach by the end-users (researchers/scientists in the laboratory) for predicting the oil recovery. Hence, in the present study, 12 different ML models/algorithms are used in the PIML approach to determine its accuracy in predicting the output oil recovery. The 12 different ML models/algorithms that were used in the present study are K-Nearest Neighbours (KNN), Decision Trees, Lasso, Ridge, Linear Regression, Random Forests, ADA Boost Regression, Gradient Boosting, Gaussian Process Regression, Polynomial Regression, Support Vector Regression (SVR) and neural networks.

For all these 12 ML models adopted in the PIML approach, the input parameter data was normalised, and was subsequently split into training data sets and test data sets in the ratio of 7:3 for training and testing of the ML model used. k-fold cross validation was performed on the training set by setting k = 10, and the best model is then evaluated on the test data set. In the present work, all the 12 ML models were trained using training data sets, and its prediction performance were compared based on 3 metrics, namely, Root Mean Square Error {*RMSE*; eq. (9)}, Coefficient of Determination {R^2; eq. (10)} and Explained Variance Score {*EVS*; eq. (11)}.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

$$EVS = 1 - \frac{Var(y_i - \hat{y}_i)}{Var(y_i)} \tag{11}$$

In the Eqs. (9 - 11), where:y_i represents the actual % oil recovery determined from physics-based model; $\hat{y}_i$ represents the predicted value of % oil recovery determined from ML model; $\bar{y}$ represents the mean value of y_i; *n* represents the number of samples; and *Var* represents the variance. *RMSE* is a measure of accuracy, and lower values indicate better fit of data. R^2 and *EVS* measures proportion to which a mathematical model accounts for variation of a given data set. The ML model having values of R^2 and *EVS* closer to 1 is the most accurate and suitable model that shall be used to predict the % oil recovery for in-situ MEOR application.

3. Results and Discussion

3.1 Validation of physics based microbial kinetic model

The validity of the physics-based microbial kinetic model that is used in the present study is verified by comparing the present numerical model results with the experimental data. From Fig. 3(a - c), it is observed that the present model results (microbial, nutrient and bio-surfactant concentrations *w.r.t* time) is in good agreement with the experimental data. As the present adopted model is validated, it is subsequently used to generate large datasets.

[Figure 3]

3.2 Quantifying the influence of input parameters on output oil recovery

[Figure 4]

Figure 4a shows the PCC and SPCC values in a matrix form that represents the interaction (or) association strength between any two parameters involved in the MEOR process. Fig. 4b specifically presents the PCC and SPCC values (i.e., interaction strength) between all the input parameters with the output oil recovery parameter. Results from Fig. 4a and Fig. 4b reveals that the input parameters, $Y_{PS}, u, S_{ori}, \mu_w,\ X_i,$ $A_i, S_i, U_{max},\ S_{wir}\ and\ T_r$ are directly proportional to the oil recovery, while the input parameters K_{XS}, Y_{XS} *and initial IFT* are inversely proportional to the oil recovery. These results are consistent with the reality, which validates the results shown in Fig. 4.

It is observed that among all these input parameters, the parameter, Y_{PS} had strongly associated with the output oil recovery while compared to all the other input parameters. This illustrates that the yield value of biosurfactants *w.r.t* sucrose (Y_{PS}) is the dominant parameter that significantly influences the output oil recovery. Moreover, it is also inferred that the oil recovery performance of MEOR process increases with increase in Y_{PS} value, which means that with the increase in utilization of nutrients for biosurfactant production, the oil recovery increases. This obtained result corroborates with the earlier results of Sivasankar and Suresh Kumar, 2019, in which, it is reported that Y_{PS} parameter significantly influences the oil recovery compared to other kinetic parameters. From Fig. 4a and Fig. 4b, it is also observed that among the negatively correlated input parameters (*i.e*., parameters that are inversely proportional to the oil recovery), Y_{XS} is the input parameter that is strongly associated with the output oil recovery. This illustrates that lower the value of Y_{XS} (*i.e*., less nutrient is utilized for the growth of microbes), higher is the oil recovery. Hence, the study reveals that the higher oil recovery is attained by selecting a microbe-nutrient pair that have higher value of Y_{PS} and lower value of Y_{XS}. Based on these observations made on Y_{PS} and Y_{XS} values, it shall be finally correlated that the ratio between Y_{PS} and Y_{XS} (*i.e*., Y_{PS}/Y_{XS}) values for a microbe-nutrient pair needs to be higher to achieve better oil recovery. Thus, based on the determination of PCC and SPCC values, it is concluded that: (a) Y_{PS} and Y_{XS} are the two input parameters that significantly influences the output oil recovery; and (b) oil recovery could be maximized by selecting a microbe-nutrient pair having higher Y_{PS}/Y_{XS} value at initial laboratory investigation stage itself.

3.3 Application of PIML approach for identifying critical input microbial kinetic, operational and reservoir parameters influencing the output oil recovery

[Figure 5]

Figure 5a presents the Relative Importance (RI) score or relative strength of all the 13 input parameters on influencing the output oil recovery. The RI score was determined by performing feature importance study. As all the 13 input parameters involved in the feature importance study were selected (or) sourced from the physics-based model (Eqs. 1- 8), hence, all these input parameters are relevant and have some influence on deciding the output oil recovery. This is evident from Fig. 5a, which shows that each of the 13 input parameters have a non-zero RI score. Thus, in the present PIML approach, all the 13 input parameters are considered for the training, testing, and implementation of all ML algorithms (models) for predicting the output oil recovery.

It is observed from Fig. 5a that among all the input parameters, Y_{PS} has the highest RI score of 0.168, hence, it is the most critical input which significantly influences the output oil recovery. In order to exclusively understand the relative importance of microbial, operational and reservoir parameters on deciding the output oil recovery, correspondingly, Figs. 5b, 5c and 5d are plotted. It is understood from Fig. 5b that among the input parameters that are related to microbes and nutrients (i.e., Y_{PS}, Y_{XS}, S_i, A_i, U_{max}, X_i, K_{XS}), Y_{PS} and Y_{XS} are relatively the most influential input parameters with RI score of 0.168 and 0.1, respectively. While, K_{XS} is relatively the less significant input kinetic parameter on deciding the percent of output oil recovery with RI score of 0.014. It is also observed from Fig. 5b that compared to injection concentration of microbes, the injection concentration of nutrients (both, carbon and nitrogen source) into the reservoir has relatively higher impact on deciding the output oil recovery. This implies that for maximizing the oil recovery, continuous supply of nutrients to the microbes need to be ensured for microbes to undergo metabolic activity within the reservoir (*i.e.*, to produce biosurfactants) and recover the oil.

Fig. 5c shows the relative importance of operational parameters (u, μ_w, T_r) on influencing the output oil recovery. It is observed from Fig. 5c that though all operational parameters influence the output oil recovery, fluid velocity (u) is the input operational parameter that influences the output oil recovery relatively more, and closely followed by the viscosity of injection water (μ_w) parameter. This obtained results are in accordance with the physics-based concept of Capillary Number, in which, the viscous force (*i.e*, product of u and μ_w) must be higher for achieving higher oil recovery. Thus, the results from Fig. 5c implies that more oil could be recovered during field implementation of in-situ MEOR technique by optimizing the injection velocity of the microbial slug (*i.e*, mixture of microbes, nutrients and water) and by increasing the water viscosity using biopolymer producing microbes during in-situ MEOR application.

Figure 5d presents the relative importance scores of different parameters (*i.e.*, $S_{ori}, S_{wir}, \sigma_i$) related to the fluids present within the reservoir. By correlating the results from Fig. 5d and from Fig. 4b, it is inferred that among the fluid parameters, the initial residual oil saturation parameter (S_{ori}) is the most significant parameter influencing the oil recovery, and the oil recovery will be higher in reservoirs that has higher value of S_{ori}. This finding is in good agreement with the earlier physics-based simulation studies (Sivasankar et al., 2016) which states that the oil recovery performance increases with the increase in initial residual oil saturation. Hence, based on this finding from Fig. 5d it is suggested that the oil recovery performance of MOER technique could be improved if the MEOR technique is implemented at an earlier stage of oil production (i.e., along with secondary recovery stage), during which the oil saturation will be relatively higher compared to the later stage (i.e., at tertiary recovery stage).

As the results presented in Fig. 5 (a – d) validates with the physics-based model results, it has been affirmed that the physics has been infused into the ML model in the present PIML approach. Hence the results obtained from this PIML approach can be used to draw physical insights, based on which, suitable strategies can be developed for maximizing the oil recovery.

3.4 Application of PIML approach to screen suitability of in-situ MEOR technique and to identify suitable microbe-nutrient for in-situ MEOR implementation

The RI score of each input parameter presented in Fig. 5a also correspondingly represents the weightage factor of each input parameter. Based on this weightage factor, the selection score of in-situ MEOR technique is calculated. This selection score helps in initial screening of in-situ MEOR technique among other EOR techniques for field implementation. The EOR technique that possess the highest selection score will be considered further for field implementation. Earlier, the selection sore for in-situ MEOR technique was calculated based on the reservoir fluid and rock properties, and neglected the consideration of microbial kinetic and operational parameters, which may mislead the entire EOR screening process for field implementation. However, with the PIML approach adopted in the present work, the selection score for in-situ MEOR technique is calculated by including both microbial kinetic parameters (Y_{PS}, Y_{XS}, K_{XS}) and operational parameters along with reservoir fluid and rock parameters. Thus, the present work would enhance the accuracy in screening of in-situ MEOR technique, which subsequently would help the end users to make better decision on selecting a suitable EOR for field implementation. The selection score for an EOR technique is calculated by using the formula, $Selection\ Score = \sum_{i=1}^{n} w_i a_i$. Here, i represents the input parameters; n represents the total number of input parameters; a_i represents the accuracy factor of input parameter, i, and its value varies between 0 and 1. Accuracy factor value represents the measure of closeness of that input parameter value with the most favourable value range; w_i represents the weightage function of the input parameter, i, and its value varies between 0 and 1. This weightage factor represents the relative importance of that input parameter influencing the output parameter.

In calculation of selection score for present in-situ MEOR technique, the weightage factor, w_i, of different input parameters, i, are same as the RI score of different input parameters as shown in Fig. 5a. Hence, based on the determined weightage factor (*i.e.*, RI score) for all the 13 input parameters, the selection score for the in-situ MEOR technique shall be calculated by using Eq. (12).

$$Selection\ Score = 0.168\ a_{Yps} + 0.146\ a_{Sori} + 0.14\ a_{u} + 0.125\ a_{\mu w} + 0.114\ a_{Ai} + 0.1\ a_{Yxs} + 0.08\ a_{Si} + 0.047\ a_{Umax} + 0.02\ a_{Tr} + 0.018\ a_{Swir} + 0.016\ a_{Xi} + 0.014\ a_{Kxs} + 0.012\ a_{\sigma i} \quad (12)$$

The value of accuracy factor values of each input parameter (a_i) are determined from lab experiments. The value of a_i varies case-to-case basis, and its value depends on the nature of reservoir and microbe-nutrient pair used and the operational conditions adopted. Upon calculation of a_i from initial experiments, the selection score for in-situ MEOR technique shall be quickly calculated using Eq. (12), which will subsequently help to screen the suitability of in-situ MEOR technique among other EOR techniques for field implementation at the initial laboratory investigation itself. In addition to it, the selection score presented in Eq. (12) also helps to identify the suitable microbe-nutrient combination among several available combinations for attaining better oil recovery at the initial laboratory investigation itself. The microbe-nutrient combination that have highest selection score value will recover relatively more oil from the reservoir for a given reservoir and operational conditions. Thus, it is concluded that the RI score determined from feature importance study in present PIML approach will: (a) help to screen the suitability of in-situ MEOR technique among other EOR techniques for field implementation; and also (b) helps to screen the suitable microbe-nutrient combination for successful implementation of in-situ MEOR technique in the field at the initial laboratory investigation itself.

3.5 Application of different ML algorithms in the PIML approach to predict the oil recovery performance of in-situ MEOR technique

[Figure 6]

Figure 6 shows the oil recovery (in %) predicted by different ML models used in the PIML approach against the benchmark (actual) results which are obtained from physics-based models. The most accurate ML model with better prediction capability will have the scatter plot points lying closer to the line equation $y' = x$ (here, y' and x are benchmark and predicted values, respectively), and correspondingly will have R^2 and *RMSE* value closer to 1 and 0, respectively. While, for the ML model with least accuracy, the scatter plot points are spread widely from the line equation $y' = x$, and it will also have relatively lower R^2 and relatively higher *RMSE* value. The R^2, *RMSE* and *EVS* values of all the 12 ML algorithms that were used in the present PIML approach study is presented in Table 2.

[Table 2]

Based on the results presented in Fig. 6 and Tab. 2, it is inferred that among all the 12 ML algorithms/models that are used in the present PIML approach, Neural Networks ML model had performed better in predicting the output oil recovery (R^2 = 0.9873, *RMSE* = 0.7145). The neural networks ML algorithm/model outperforms other ML algorithms in prediction because it can implicitly detect complex non-linear relationships between dependent and independent variables, and it also have the ability to detect all possible interactions between the input variables. Followed by the neural network model, it is observed that the Support Vector Regression (SVR) is the second-best ML model that can better predict the oil recovery ($R^2 = 0.9644$; *RMSE* = 1.184). The main advantage of SVR model is that it is less susceptible to outliers than other data-driven models but it's harder to manually tune hyperparameters. Next to SVR algorithm, it is found that the 4th degree Polynomial Regression model had predicted the oil recovery better ($R^2 = 0.963$; *RMSE* = 1.26) as it has the ability to better map the non-linear relationship between the input and output variables. Amongst all the 12 ML models that were used in the present PIML approach for oil recovery prediction, it is found that K-Nearest Neighbours ML model is the least accurate model ($R^2 = 0.3698$; *RMSE* = 3.369). Thus, from the present study, it is concluded that to predict the oil recovery performance of in-situ MEOR technique at initial lab stage, the Neural Network is the best ML algorithm that need to be used in the PIML approach.

3.5 Case study on application of PIML approach for screening of suitable microbe-nutrient combination for in-situ MEOR implementation

To illustrate the application of present PIML approach on screening of suitable microbe-nutrient combination, a case study using synthetic data has been carried out.

[Table 3]

Table. 3 presents the microbial kinetic parameters for 4 different combinations of microbes and nutrients, and rest all other parameters are kept constants. By feeding the inputs through the trained neural network ML algorithm, the output oil recovery is calculated and presented in the last column of Tab 3. It is inferred from Tab.3 that among all the available combinations, the combination 4 shows highest oil recovery, hence that corresponding microbe-nutrient pair can be used for field implementation.

4. Conclusions

Physics-Informed Machine Learning (PIML) approach is adopted to investigate the performance of in-situ MEOR technique from limited experimental data, which is difficult with conventional experimental and modelling approaches. Neural network ML model used in the PIML approach had more accurately predicted the oil recovery. Y_{PS}, flow velocity and initial oil saturation (S_{ori}) are correspondingly the most influential microbial kinetic, operational and reservoir parameter. Higher oil recovery is achieved by selecting a microbe-nutrient-reservoir pair having higher Y_{PS}/Y_{XS} and S_{ori} values. This PIML approach helps to screen/identify suitable microbe-nutrient-reservoir pair at initial laboratory stage itself, ensuring its success during the field implementation.

Funding Source: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Ansah, E.O., Thanh, H.V., Sugai, Y., Nguele, R., Sasaki, K., 2020. Microbe-induced fluid viscosity variation: field-scale simulation, sensitivity and geological uncertainty. J Petrol. Explor. Prod. Technol. 10, 1983–2003.

2. Cruz, I.A., Chuenchart, W., Long, F., Surendra, K.C., Andrade, L.R.S., Bilal, M., Figueiredo, R.T., Khanal, S.K., Ferreira, L.F.R., 2021. Application of machine learning in anaerobic digestion: Perspectives and challenges. Bioresour. Technol. 126433.

3. Joshi, S.J., Al-Wahaibi, Y.M., Al-Bahry, S.N., Elshafie, A.E., Al-Bemani, A.S., Al-Bahry, A., Al-Mandhari, M.S., 2016. Production, characterization, and application of bacillus licheniformis W16 biosurfactant in enhancing oil recovery. Front. Microbiol. 7, 1853.

4. Jeong, M.S., Lee, Y.W., Lee, H.S., Lee, K.S., 2021. Simulation-Based Optimization of Microbial Enhanced Oil Recovery with a Model Integrating Temperature, Pressure, and Salinity Effects. Energies, 14, 1131.

5. Jeong, M.S., Cho, J., Lee, K.S., 2022. Systematic modelling incorporating temperature, pressure, and salinity effects on in-situ microbial selective plugging for enhanced oil recovery in a multi-layered system, Biochem. Eng. J. 177, 108260.

6. Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nat. Rev. Phys. 4, 422-440.

7. Keprate, A., Ratnayake, R.M.C., 2019. Data Mining for Estimating Fatigue Strength Based on Composition and Process Parameters. Proc. ASME 2019 38th Int. Con. on Ocean, Offshore and Arctic Eng. Vol 4: Materials Technology.

8. Liu, H., Zhang, J., Liang, F., Temizel, C., Basri, M.A., Mesdour, R., 2021. Incorporation of physics into machine learning for production prediction from

unconventional reservoirs: a brief review of the gray-box approach. SPE Res. Eval. Eng. 24, 847–858.

9. Markande, A.N., Patel, D., Varjani, S.J., 2021. A review on biosurfactants: properties, applications and current developments. Bioresour. Technol. 330, 124963.

10. Nielsen, S.M., Nesterov, I., Shapiro, A.A., 2016. Microbial enhanced oil recovery-a modeling study of the potential of spore-forming bacteria. Comput. Geosci. 20, 580.

11. Nikolova, C., Gutierrez, T., 2020. Use of microorganisms in the recovery of oil from recalcitrant oil reservoirs: current state of knowledge, technological advances and future prespective. Front. Microbiol. 10, 2996.

12. Sivasankar, P., Kanna, R., Kumar, G.S., Gummadi, S.N., 2016. Numerical modelling of biophysicochemical effects on multispecies reactive transport in porous media involving *Pseudomonas putida* for potential microbial enhanced oil recovery application. Bioresour. Technol. 211, 348-359.

13. Sivasankar, P., Kumar, G.S., 2017. Influence of pH on dynamics of microbial enhanced oil recovery processes using biosurfactant produced *Pseudomonas putida*: Mathematical modelling and numerical simulation. Bioresour. Technol. 224, 498-508.

14. Shibulal, B., Al-Bahry, S.N., Al-Wahaibi, Y.M., Elshafie, A.E., Al-Bemani, A.S., Joshi, S.J., 2018. Microbial-Enhanced Heavy Oil Recovery under Laboratory Conditions by *Bacillus firmus* BG4 and *Bacillus halodurans* BG5 Isolated from Heavy Oil Fields. Colloids Interfaces. 2, 1.

15. Sivasankar, P., Kumar, G.S., 2019. Influence of bio-clogging induced formation damage on performance of microbial enhanced oil recovery processes. Fuel. 236, 109.

16. Thanh, H.V., Sugai, Y., Sasaki, K., 2020. Application of artificial neural network for predicting the performance of CO_2 enhanced oil recovery and storage in residual oil zones. Sci. Rep. 10, 18204.

17. Tang, Q., Chen, Y., Yang, H., Liu, M., Xiao, H., Wang, S., Chen. H., Naqvi, S.R., 2021. Machine learning prediction of pyrolytic gas yield and compositions with feature reduction methods: Effects of pyrolysis conditions and biomass characteristics. Bioresour. Technol. 339, 125581.

18. Varjani, S.J., Upasani, V.N., 2016. Core Flood study for enhanced oil recovery through ex-situ bioaugmentation with thermo- and halo-tolerant rhamnolipid produced by Pseudomonas aeruginosa NCIM 5514. Bioresour. Technol. 220, 175-182.

19. Varjani, S.J., Upasani, V.N., 2017. Critical review on biosurfactant analysis, purification and characterization using rhamnolipid as a model biosurfactant. Bioresour. Technol. 232, 389-397.

20. Wang, Z., Peng, X., Xia, A., Shah, A.A., Huang, Y., Zhu, X., Zhu, X., Liao, Q., 2022. The role of machine learning to boost the bioenergy and biofuels conversion. Bioresour. Technol. 343, 126099.

21. Zhang, W, Li, J., Liu, T., Leng, S., Yang, L., Peng, H., Jiang, S., Zhou, W., Leng, L., Li, H., 2021. Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. Bioresour. Technol. 342, 126011.

Figures Caption

1. Procedure of PIML approach followed in the present study to predict oil recovery by in-situ MEOR process

2. Frequency distribution of input and output data that are generated and used for training and testing of ML algorithms in the PIML approach

3. Validation of present microbial kinetic model results with measured experimental data for (a) variation of microbial concentration with time, (b) variation of sucrose concentration with time, (c) variation of biosurfactant concentration with time

4. (a) Pearson correlation and Spearman correlation coefficient matrix for microbe, operational, reservoir and % oil recovery data, (b) Correlation coefficient values for input microbe, operational and reservoir data towards output % oil recovery

5. (a) Relative Importance (RI) score of all input parameters, (b) RI score of input microbial-nutrient parameters, (c) RI score of input operational parameters, (d) RI score of input reservoir parameter in predicting output % of oil recovery

6. Comparative performance of 12 different ML algorithms in predicting the oil recovery against the actual % of oil recovery for in-situ MEOR application.

Tables Caption

1. Input parameters and their corresponding value range used in the present study

2. Performance of different ML algorithms in predicting the actual % of oil recovery

3. Microbial kinetic parameters for different microbial-nutrient combinations and their corresponding oil recovery determined using PIML modelling approach.

Table 1. Input parameters and their corresponding value range used in the present study [Sivasankar et al. (2016)]]

Parameter	**Reference value**	**Range**
Y_{XS}	0.1843	0.092 - 0.276
Y_{PS}	0.078 [Sivasankar et al. (2016)]	0.03900733 - 0.116996715
K_{XS} (*g/l*)	6.86 [Sivasankar et al. (2016)]	3.430058808 - 10.28959695
U_{max} (h^{-1})	0.053 [Sivasankar et al. (2016)]	0.02650387 - 0.079495509

X_i (*g/l*)	0.1521167 [Sivasankar et al. (2016)]	0.076094593 - 0.22824074
S_i (*g/l*)	19.234 [Sivasankar et al. (2016)]	9.617601084 - 28.84936382
A_i (*g/l*)	3 [Sivasankar et al. (2016)]	1.500165456 - 4.49988547
T_r (*h*)	150	100 – 200
u_w (*m/h*)	0.0004 [Sivasankar et al. (2016)]	0.0002 – 0.0006
μ_w (Nhm^{-2})	0.001 [Sivasankar et al. (2016)]	0.0005 – 0.0015
Initial IFT (*mN/m*)	51.6 [Sivasankar et al. (2016)]	25.80405697 - 77.39695
S_{wir}	0.2	0.10000517 - 0.299989777
S_{ori}	0.4	0.20000454 - 0.599986979
Output - % oil recovery	Mean - 7.423469637 Median - 5.510703244	0.174742622 - 48.23409386

Table 2: Performance of different ML algorithms in predicting the actual % of oil recovery

Model	R^2	***RMSE***	***Explained Variance Score***
KNN	0.369897442	3.369742385	0.370031425
Decision Trees	0.408587985	4.356687642	0.408683171
Lasso	0.510880307	3.697724477	0.51138261
Ridge	0.512474691	3.697351147	0.512976983

Linear Regression	0.51299017	3.697264614	0.513491055
Random Forests	0.639724278	2.941482946	0.63972701
ADA Boost	0.639746354	2.811959461	0.639980378
Gradient Boosting	0.896025654	1.851540808	0.896025658
Gaussian Process	0.951787039	1.353896887	0.951802595
Polynomial (4)	0.963022723	1.26029449	0.963039744
SVR	0.964436929	1.184682456	0.964596291
Neural Network	0.987349995	0.714597767	0.987656577

Table 3: Microbial kinetic parameters for different microbial-nutrient combinations and their corresponding oil recovery determined using PIML modelling approach

Combinations	$\boldsymbol{Y_{XS}}$	$\boldsymbol{Y_{PS}}$	$\boldsymbol{K_{XS}}$	$\boldsymbol{U_{max}}$	**Output Oil Recovery, %**
1	0.098734	0.067978	4.077158	0.068247	6.3529167
2	0.148067	0.081408	4.763728	0.056552	4.531987

3	0.169445	0.091877	4.923773	0.053963	4.549041
4	0.121985	0.095722	8.262717	0.038336	7.3623314

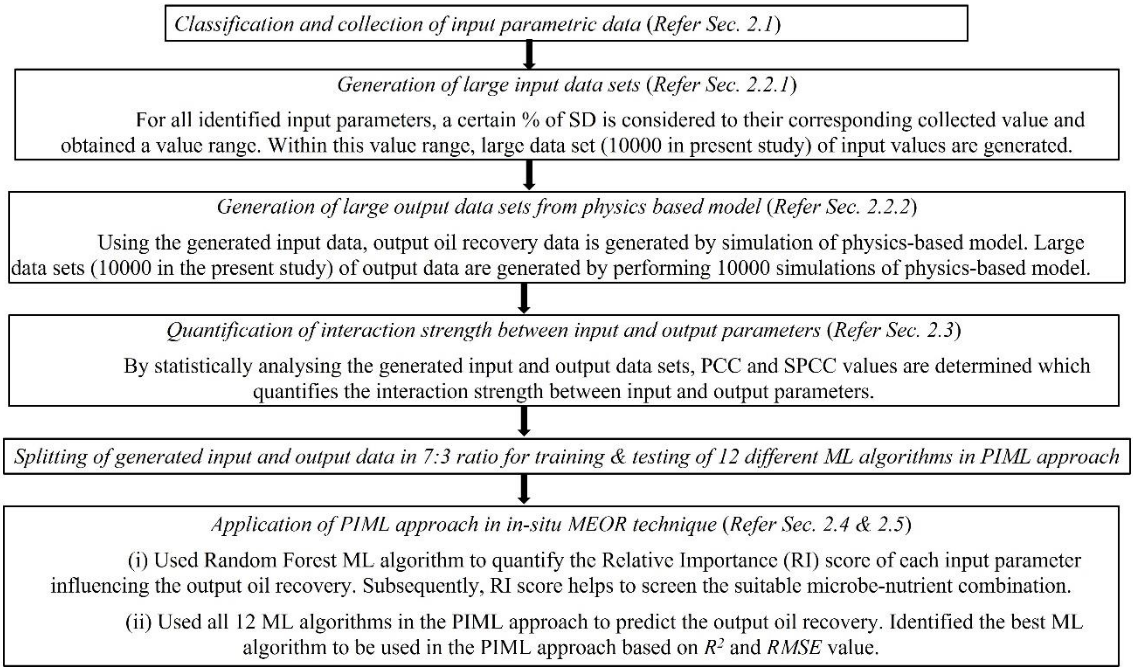


Figure 1: Procedure of PIML approach followed in the present study to predict oil recovery by in-situ MEOR process (NIKHIL WILL CHANGE IT)

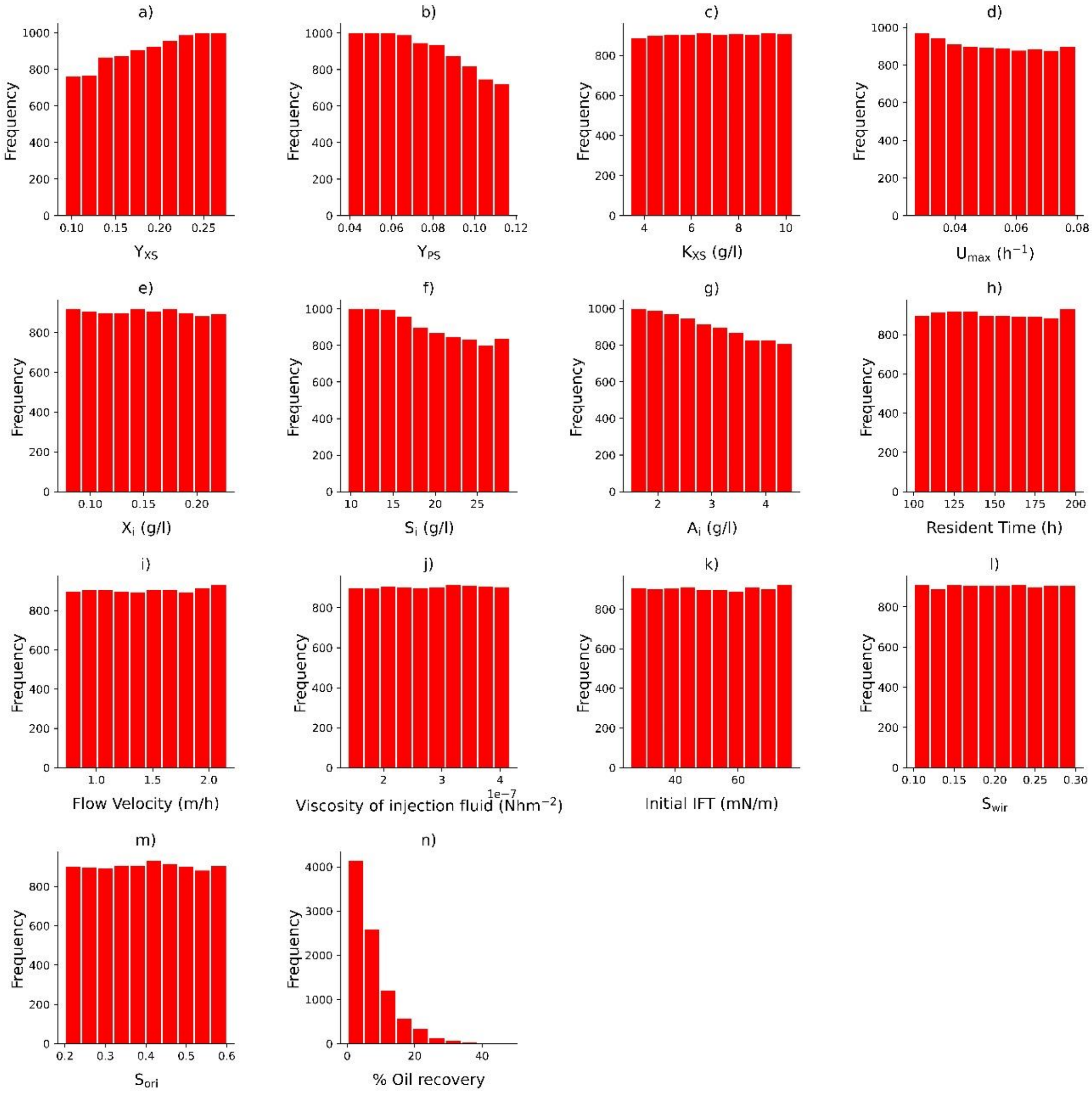


Figure 2: HIstogram of input and output data that are generated and used for training and testing of ML algorithms in the PIML approach.

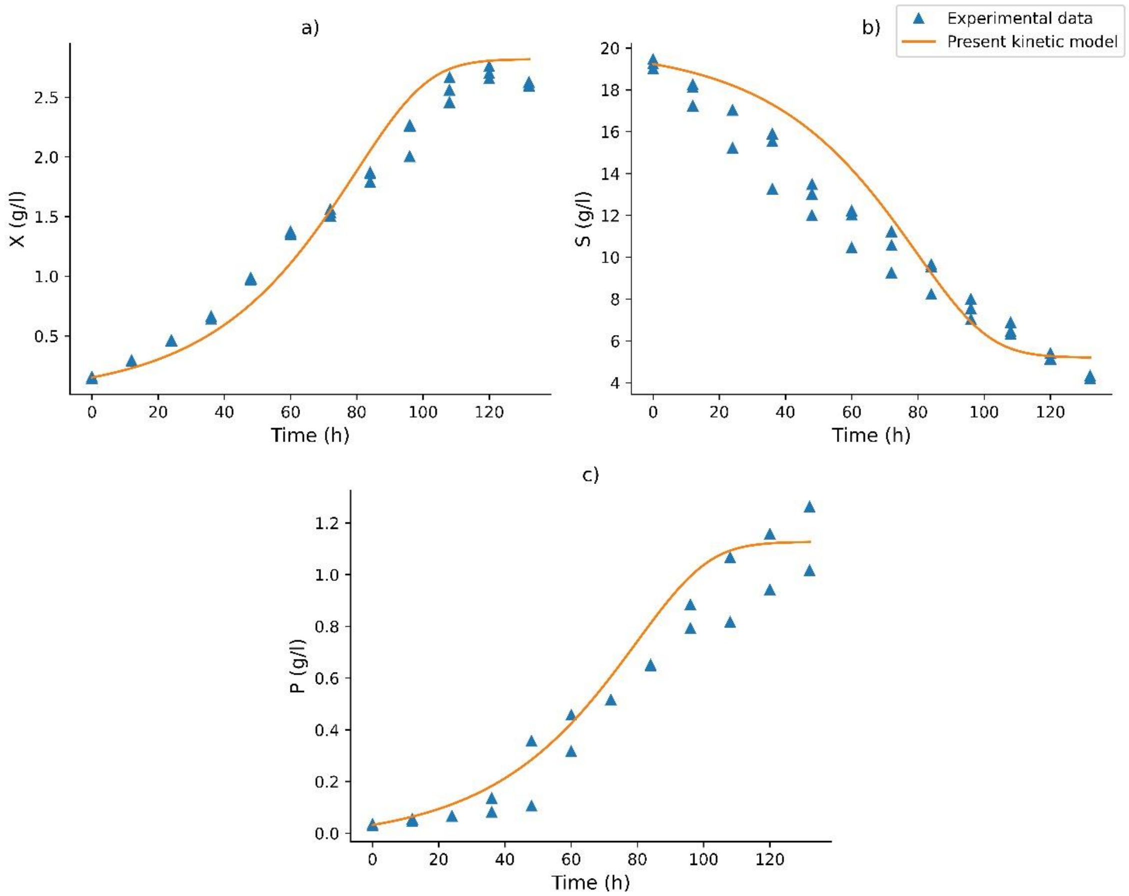
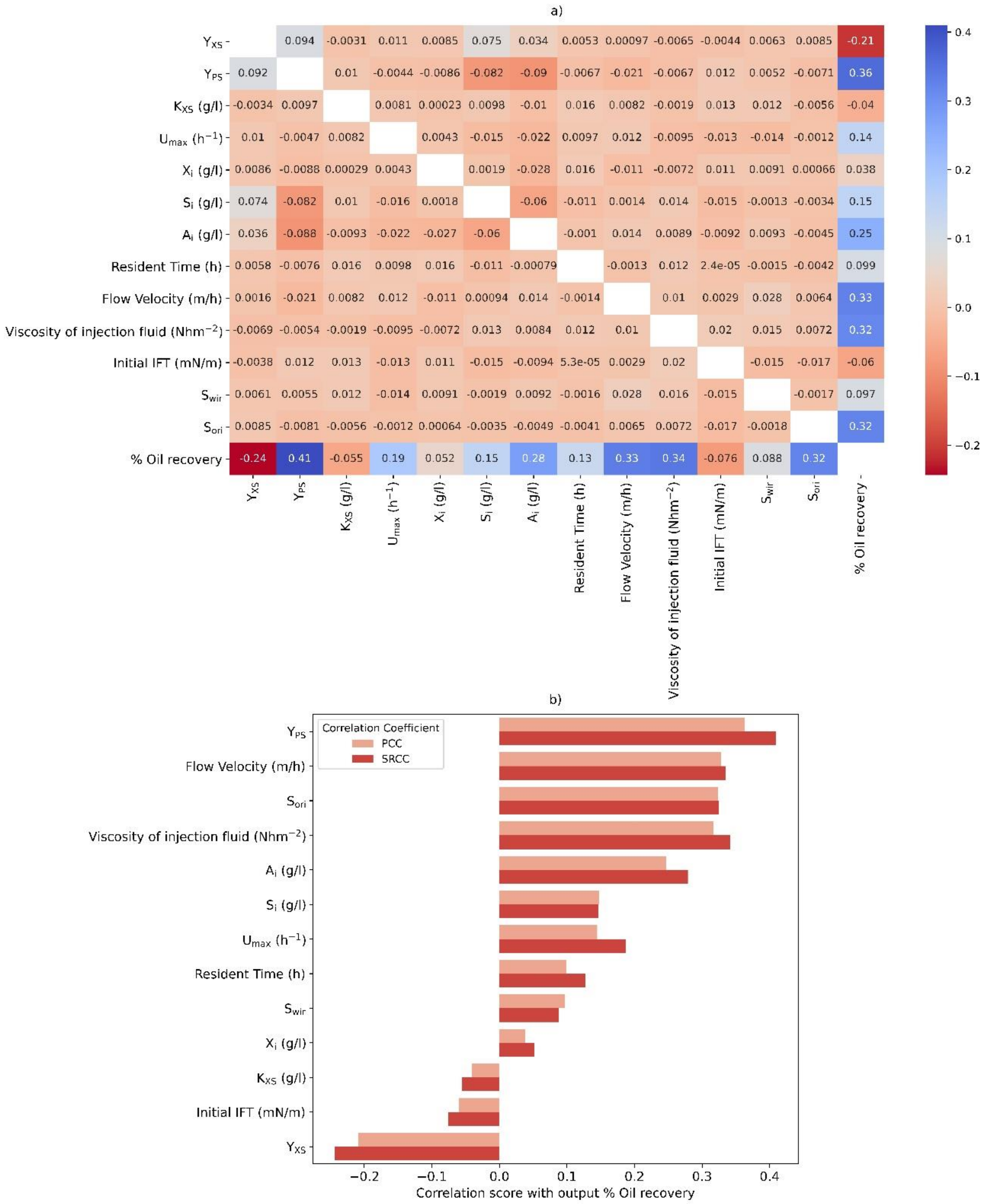


Figure 3: Validation of present microbial kinetic model results with measured experimental data for (a) variation of microbial concentration with time, (b) variation of sucrose concentration with time, (c) variation of biosurfactant concentration with time.



668

669 **Figure 4:** (a) Pearson correlation and Spearman correlation coefficient matrix for

670 microbe, operational, reservoir and % oil recovery data, (b) Correlation coefficient

671 values for input microbe, operational and reservoir data towards output % oil recovery.

672

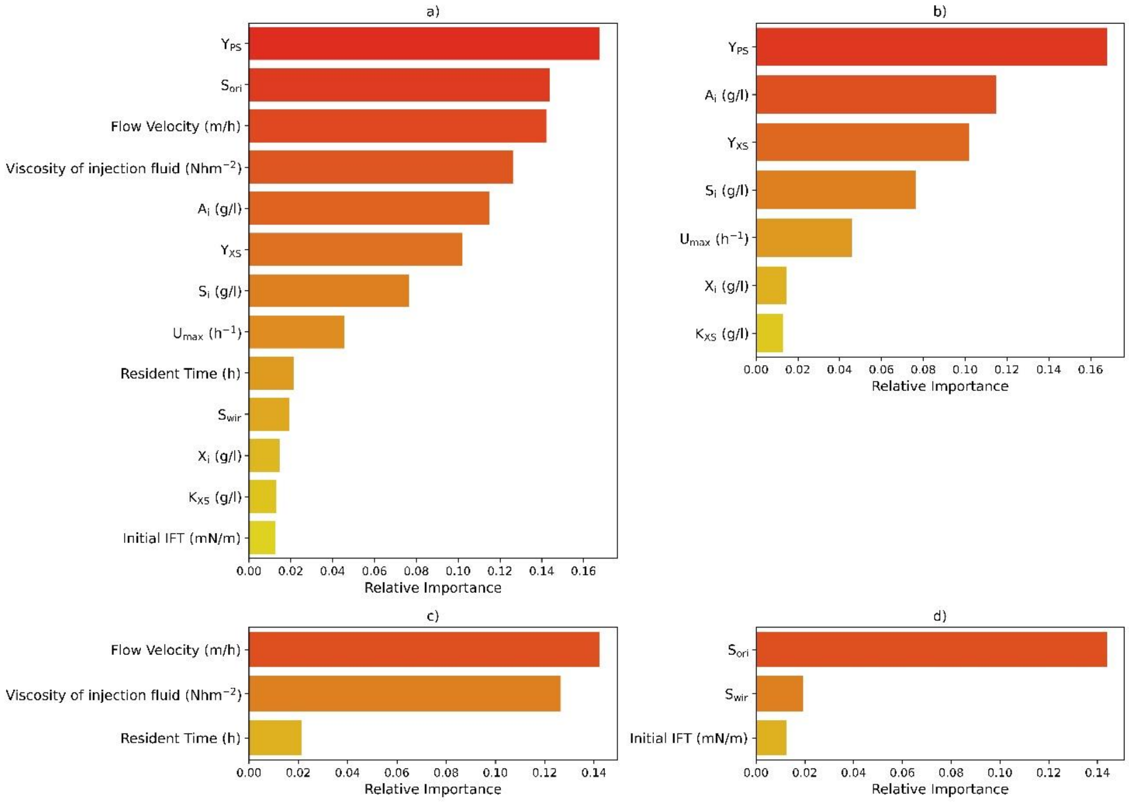


Figure 5: (a) Relative Importance (RI) score of all input parameters, (b) RI score of input microbial-nutrient parameters, (c) RI score of input operational parameters, (d) RI score of input reservoir parameter in predicting output % of oil recovery.

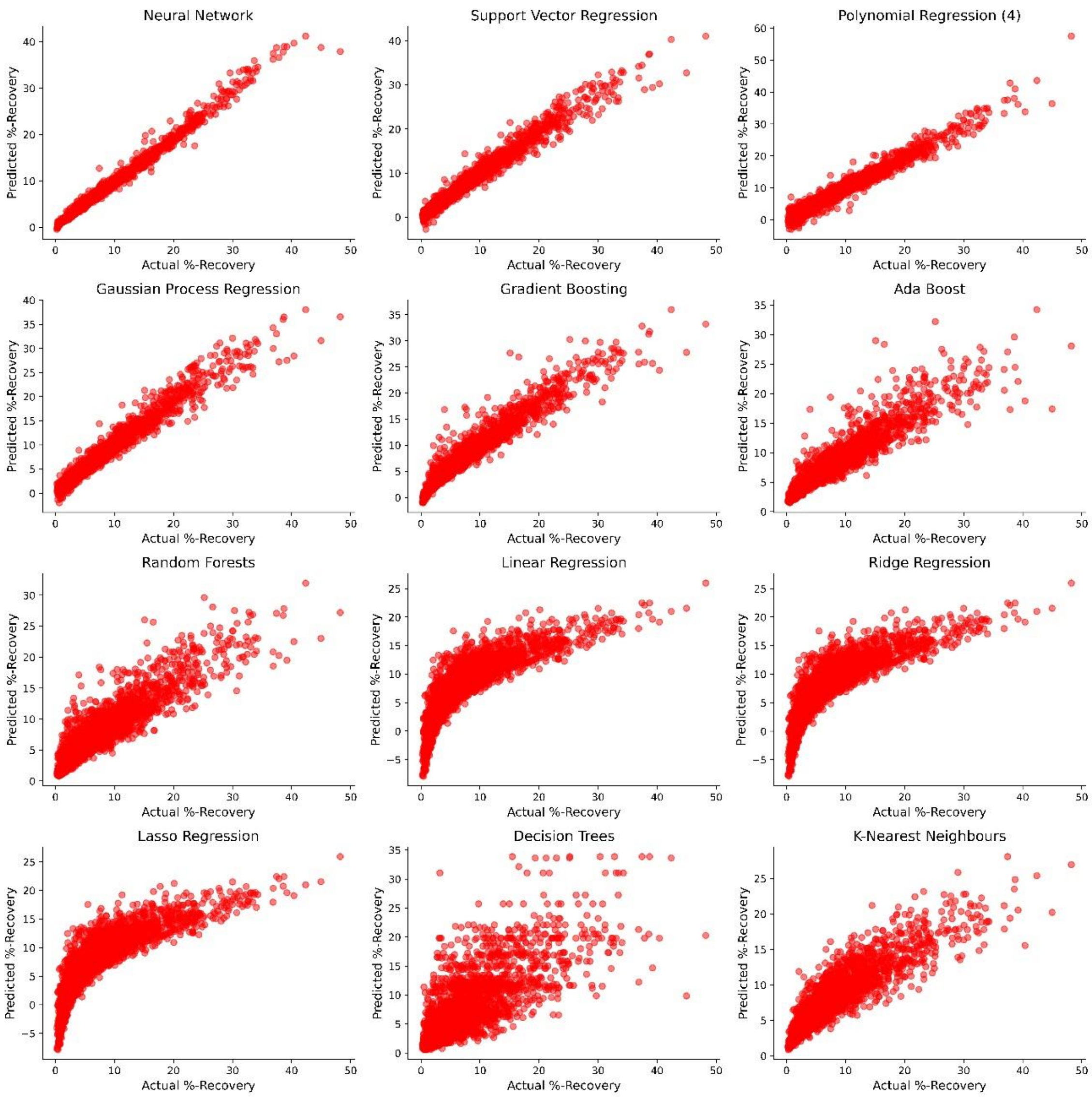


Figure 6: Comparative performance of 12 different ML algorithms in predicting the oil recovery against the actual % of oil recovery for in-situ MEOR application.