**ABUAD Journal of Engineering Research and Development (AJERD)**
**ISSN (online): 2645-2685; ISSN (print): 2756-6811**

Volume 8, Issue 1, 292-306

# Prediction of Reservoir Flow Capacity in Sandstone Formations: A Comparative Analysis of Machine Learning Models

Micheal Ayodeji OGUNDERO[1], Taiwo ADELAKIN[1], Kehinde OROLU[1], Isaac Femi JOHNSON[1], Theophilus Akinfenwa FASHANU[1], Kingsley E. ABHULIMEN[2]

[1]*Department of Systems Engineering, University of Lagos, Akoka, Lagos, Nigeria*
mogundero@unilag.edu.ng/reachtaye@gmail.com/korolu@unilag.edu.ng/femijohnsonengltd@gmail.com/tfashanu@unilag.edu.ng

[2]*Department of Chemical Engineering, University of Lagos, Akoka, Lagos, Nigeria*
kabhulimen@unilag.edu.ng

*Corresponding Author:* mogundero@unilag.edu.ng, +2348087365099
*Date Submitted:* 05/12/2024
*Date Accepted:* 18/04/2025
*Date Published:* 24/04/2025

***Abstract:*** *Sand production is one of the major challenges in the oil and gas industry, impacting the operational integrity and economic efficiency of oil extraction activities. This study focuses on predicting Reservoir Flow Capacity (RFC) in sandstone formations by analyzing geological and petrophysical properties critical to reservoir performance and mechanical stability. It also identified key factors that impact the mechanical stability of formations during production. Given a large number of input variables that enclose geological and environmental factors, the study set the correlation of these conditions to provide profound analysis and reveal profound patterns within the data. With the following supervised machine learning algorithms: Random Forest, Artificial Neural Network (ANN) and Support Vector Regression (SVR); the study modeled RFC. The algorithms were selected for their ability to model complex relationships in reservoir characterization, with Random Forest excelling in high-dimensional data handling, ANN in pattern learning, and SVR in regression-based predictions. Model evaluation using R-Squared metrics showed that the Random Forest model possesses a good level of accuracy of 0.9573 in predicting the RFC, compared to the ANN and SVR model which had R-Squared values of 0.9390 and 0.7294 respectively. The SVR model had large variations from the actual values and hence was not very useful for our predictions. Further analysis using the developed machine learning models revealed that geological formation thickness, reservoir thickness, and permeability are the most critical parameters influencing reservoir flow capacity and overall rock stability.*

***Keywords:*** *Sand Production, Support Vector Regression, Artificial Neural Network, Random Forest, Geological Formation*

## 1. INTRODUCTION

Crude oil drilling is a complicated exercise designed for extracting the oil found below the surface of the earth, an important activity in the global economy. Hence this process has various stages right from the selection of the site, drilling and finally production of crude oil. The drilling process can be categorized into three stages, namely: Site Selection or Preparation, Drilling Process and Well Completion and Production. The process starts by conducting geological exploration in order to establish areas which could possibly contain oil reserves. Exploration of crude oil entails making a hole through the surface of the earth to the subsurface with depths extending up to few thousands of feet. After reaching the target depth, the well is completed to allow for the safe and efficient production of oil.

Among the various challenges encountered during oil production like: produced water management, carbon capture and sequestration, in-situ molecular manipulation, etc [1]. As shown in Figure 1, it is capable of causing problems to equipment. For example, sand erosion of downhole and surface equipment and sand accumulation on surface and Sand Disposal. The management of sand production stands out due to its significant impact on operational efficiency, equipment integrity, and overall safety [2]. Sand management in oil production is a critical discipline that aims to predict, control, and mitigate the production of sand, ensuring the economic viability and longevity of oil wells [3].

The important decision-making problems in offshore drilling for crude oil is that of allocating various target locations in the field to drill, to available drilling rigs as well as selecting appropriate control measures [4]. Selecting an optimum sand control strategy depends on various factors such as reservoir characteristics, available service infrastructure, production rates, production strategy, and the extent of skin damage. The challenge of sand production in crude oil drilling necessitates a multidisciplinary approach, combining geological understanding, engineering design, and operational management to minimize its impacts and ensure safe, efficient, and sustainable oil production.

Figure 1: The Process of sand production in oil reservoirs [5]

The genesis of sand production is inherently linked to the physical and chemical interactions between the reservoir rock and the in-situ fluids under the dynamic conditions of oil extraction. Some attributes that influence occurrence of sand include the nature of the formation rocks, permeability, porosity and the state of the in-situ stresses typical of the reservoir. Other factors which include drilling practices, production volume besides well completion methodologies may enhance or minimize sand production. Sand control, therefore, pertains to establishing a proper evaluation of sand occurrence, and the interrelation between geology or the earth science, and engineering in protecting reservoirs for oil explorations.

Gaining information on flow capacities of reservoirs is important for maximizing the recovery of oil and relative stability of the production system. The abilities with which the reservoir rock allows fluid to flow are measured through properties such as permeability and porosity. Such geological properties are useful in determining the performance of the reservoirs in case of petroleum liquids. It is realized that sand production could depend directly on the lithology and other geological parameters of the reservoir rock [6]. Sand production comes about when formation rock strength proves to be unable to hold pressures, that is, when the induced stresses exceed the reservoir rock in situ strength [7]. Lithologic factors such as grain size, the amount of cement, and rock composition influence the mechanical strength of the rock [8]. These characteristics also affect the reservoir rock's susceptibility to erosion forces driven by production stresses [9]. Therefore, the geological and lithological features of the reservoir should be well understood by the intending policy maker. By using geotechnical analysis in conjunction with the flow dynamics of the reservoir, we can evaluate sand production risk and control its occurrence in the wellbore. Apart from optimizing the process of recovery of the oil, integrating geotechnical analysis with reservoir flow dynamics provides us with more insights. It ensures the maintenance of the structural integrity of the well and its surrounding formation.

Predicting sand occurrence during crude oil drilling is crucial for ensuring operational efficiency and mitigating potential production challenges. Cheddad, [10] utilized ANN, RFC and SVM to predict permeability in heterogeneous reservoirs. The study showed that integrating the Flow Zone Indicator rock typing technique with the machine learning algorithms effectively predicts permeability. However, this study did not directly predict reservoir flow capacity or address sand production issues. Similarly, Krishna et al., [11] compared nine different machine learning methods to predict pore pressure in Mangahewa gas field, New Zealand. Among these, the Decision Tree model performed excellently, achieving RMSE values between 0.25 and 14.71 psi. While the study provided us with valuable insights into pore pressure prediction, it also did not directly address RFC or sand production challenges. Ali et al., [12] employed unsupervised clustering with class-based ensemble machine learning to predict elastic logs in heterogeneous rocks. The methodology improved the accuracy of reservoir characterization by effectively handling complex geological data. Otmane et al., [13] developed a hybrid method of combining traditional reservoir simulation with machine learning techniques to boost reservoir prediction accuracy. This integration of the machine learning techniques addresses major issues of the traditional methods such as high computational costs.

The reviewed studies demonstrate the application of machine learning techniques in enhancing reservoir characterization and prediction accuracy. However, there remains a gap in research specifically targeting the prediction of Reservoir Flow Capacity (RFC) in sandstone formations concerning sand production challenges. Accurately predicting RFC is important for maintaining operational integrity and economic efficiency of oil extraction activities. Therefore, this study predicts Reservoir Flow Capacity (RFC) in sandstone formations by evaluating and comparing the performance of three distinct machine learning algorithms – Random Forest (RF), Artificial Neural Networks (ANN), and Support Vector Regression (SVR). It also identifies key geological and petro physical factors influencing RFC by using feature importance analysis. It focuses on optimizing reservoir performance and understanding the implications of geological parameters on rock stability and production efficiency.

## 2. MODEL DEVELOPMENT AND METHODOLOGY

Machine learning (ML) models represent transformative approaches to predicting the effects of the geological composition of reservoirs in the drilling of crude oil. The research design employs a comprehensive machine learning-centric approach to predicting the flow capacity within geological reservoirs.

### 2. 1 Data Collection

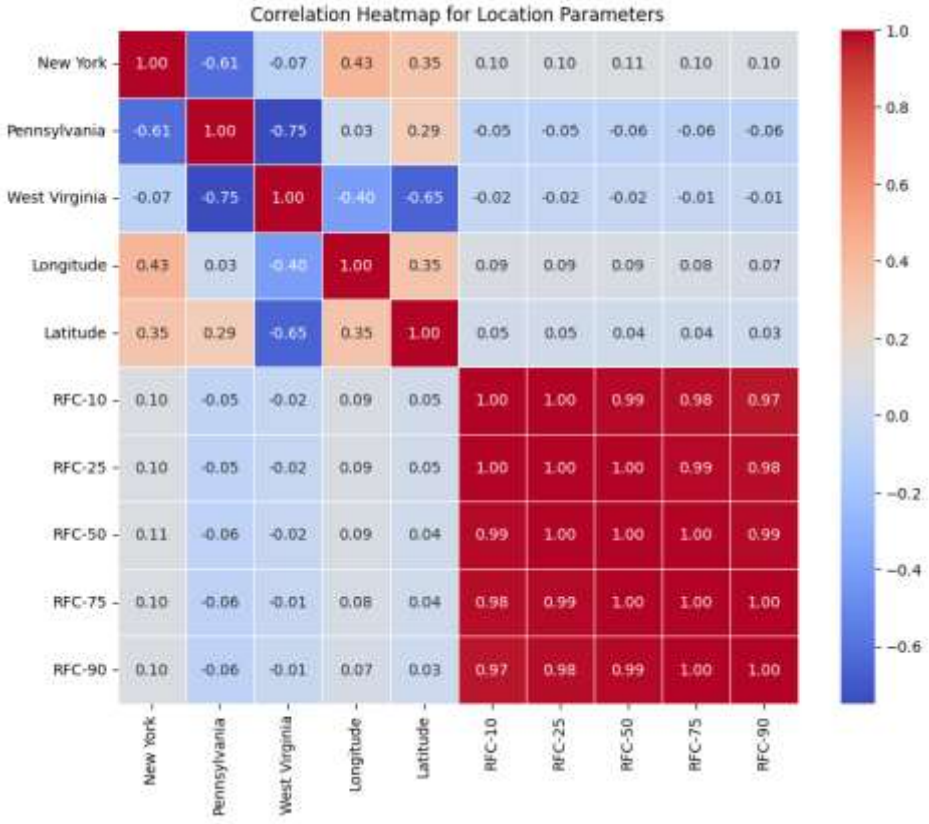
The dataset was sourced from the OpenEI's Geothermal Data Repository (GDR), [14] and contains the following features: temperature gradients, geological formations, lithology, porosity, permeability, and geospatial coordinates. This information forms the basis of our exploratory analysis and prediction modeling that will unravel the patterns for predicting the reservoir flow capacity in these formations. A statistical summary of both numerical and categorical features of the dataset is shown in Table 1 and Table 2.

### 2.2 Data Preprocessing

In order to prepare our dataset for the model training and validation, we applied important key preprocessing steps to ensure the usability and integrity of the data. We carefully assessed missing values with strategies based on the distribution of the affected feature data. The mode was used in filling missing values in the sea level elevation column because it had a skewed distribution while mean was used to populate missing reservoir temperature data because it was normally distributed [15]. Missing values in sea level elevation were imputed using the mode, while those in reservoir temperature were replaced with the mean. These approaches were selected to preserve the dataset's statistical properties while minimizing distortion in feature distribution.

Normalizing the dataset was also very important at this stage to get better results. Two different classes of data were normalized, namely: the categorical data and the numerical data. The nominal data from 'State' and ' Play type' fields were one-hot encoded as it is the effective way to feed categorical data to the ML algorithms. One-Hot Encoding is a technique used to convert categorical data into a binary format where each unique category is represented by a vector with all values set to 0 except for one position marked with a 1. This approach allows machine learning models to handle categorical variables by transforming them into a numerical format without implying any ordinal relationship. Numerical features were scaled to minimize variation thereby improving model fit and overall performance.

We then carried out feature selection by analyzing the correlation of the features with the target variable as seen in Figure 2a, 2b and 2c. Due to a very low coefficient of determination, we dropped the following geographical predictors: state codes, longitude and latitude for better model accuracy. Formation thickness, reservoir thickness and permeabilities with moderate correlation coefficients with RFC were retained while geographical features with extremely weak coefficients were removed. These are sea level elevation, porosity, reservoir temperature, reservoir depth and Lithostatic pressure.




(a)

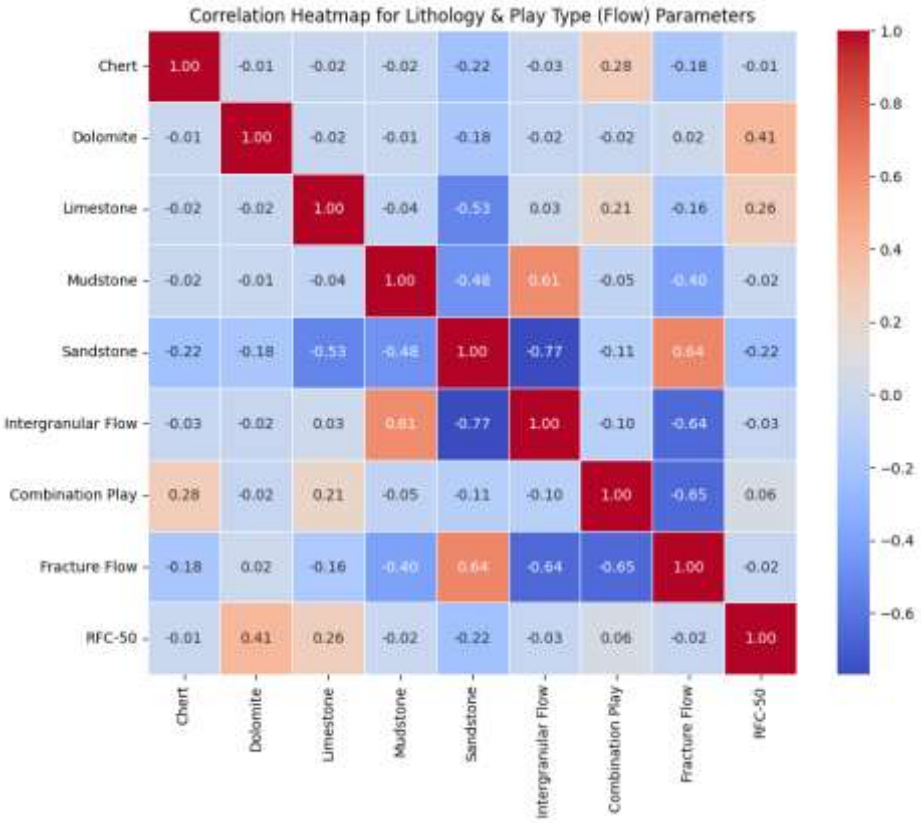
Figure 2: Correlation heatmaps: (a) location features, (b) lithological features, (c) geological features to target RFC values

Table 1: Statistical summary of numerical features of the dataset

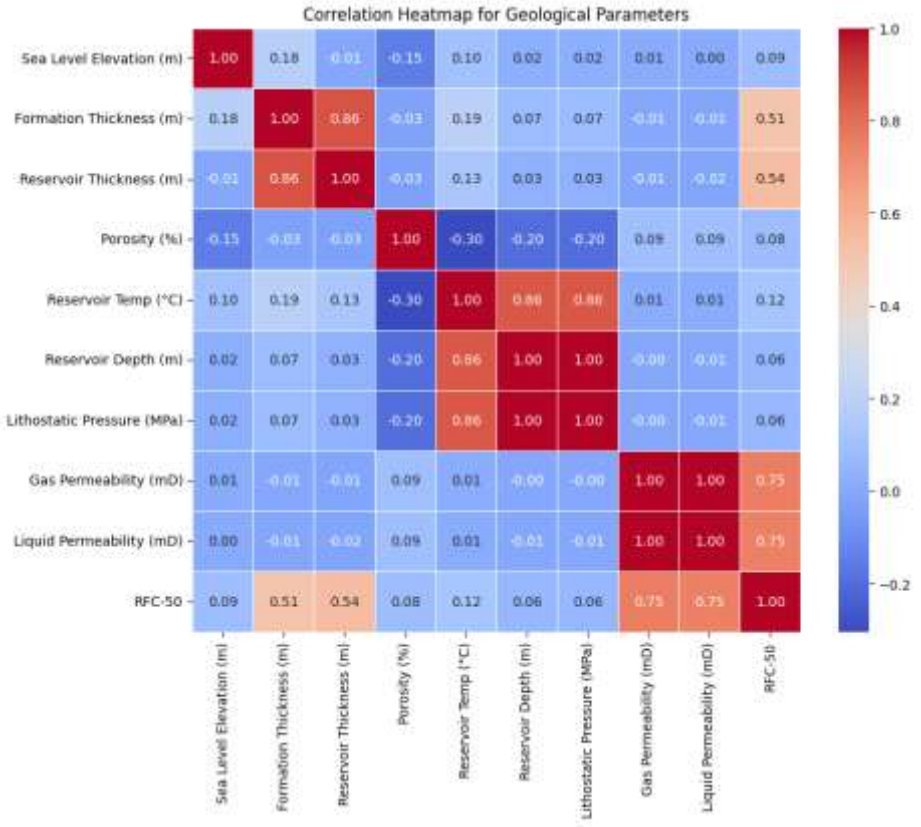
| | Latitude | Longitude | Elevation Above SeaLevel (m) | Geological Formation Thickness (m) | Reservoir Thickness (m) | Average Porocity (pct) | Reservoir Temperature (C) | Reservoir Depth (m) | Reservoir Lithostatic Pressure | Gas Permeability (mD) | Liquid Permeability (mD) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 |
| **mean** | 40.9629 | -79.5415 | 26.9208 | 24.4407 | 21.0114 | 8.5867 | 48.3874 | 1628.8207 | 41.5447 | 9.2141 | 10.7433 |
| **std** | 0.9432 | 0.9594 | 118.7086 | 53.4236 | 45.0646 | 3.2177 | 11.5487 | 415.8931 | 10.6081 | 165.3143 | 214.2580 |
| **min** | 37.2792 | -82.0044 | 0.0000 | 0.6096 | 0.6100 | 0.0000 | 17.0000 | 330.4640 | 8.4300 | 0.0010 | 0.0003 |
| **25%** | 40.2410 | -80.2062 | 0.0000 | 5.1816 | 4.5700 | 6.0000 | 41.0000 | 1353.6933 | 34.5300 | 0.1000 | 0.0542 |
| **50%** | 41.2248 | -79.7738 | 0.0000 | 15.8496 | 15.0000 | 8.0000 | 47.0000 | 1478.2800 | 37.7100 | 0.1850 | 0.0908 |
| **75%** | 41.6751 | -78.9585 | 0.0000 | 28.0416 | 27.4300 | 11.0000 | 54.0000 | 1752.6000 | 44.7000 | 2.0000 | 1.5930 |
| **max** | 43.0534 | -75.8800 | 1673.3520 | 1013.1552 | 1013.2000 | 20.0000 | 114.0000 | 3992.8800 | 101.8400 | 4152.0000 | 5384.6335 |

Table 2: Statistical summary of categorical features of the dataset

| | Location (State) | | | Lithology Type | | | | | | Play Type | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NY | PA | WV | Chert | Dolomite | Lime Stone | Mudstone | Sand Stone | Unknown | Inter-granular | Combined | Fracture | Unknown |
| **count** | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 | 1894.0000 |
| **mean** | 0.0544 | 0.8654 | 0.0803 | 0.0079 | 0.0053 | 0.0375 | 0.0375 | 0.8537 | 0.0491 | 0.0919 | 0.0929 | 0.8041 | 0.0111 |
| **std** | 226830.0000 | 0.3414 | 0.2718 | 0.0887 | 0.0725 | 0.1900 | 0.1900 | 0.3535 | 0.2161 | 0.2889 | 0.2904 | 0.3970 | 0.1047 |
| **min** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| **25%** | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 |
| **50%** | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 |
| **75%** | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0000 |
| **max** | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |




(b)




(c)

Figure 2 cont'd: Correlation heatmaps: (a) location features, (b) lithological features, (c) geological features to target RFC values

RFC_P50 was selected as the target variable because it showed a strong correlation with RFC_P10, RFC_P25, RFC_P75, and RFC_P90 as seen in Figure 4a, with a maximum deviation of only 0.02. This close relationship indicates that RFC_P50 can effectively represent all these variables. By using RFC_P50 as a single target, we minimized redundancy in the dataset and improved the overall efficiency and reliability of the model. Finally, to ensure the model's ability to generalize, the dataset was split into training and testing sets, with 80% used for training and 20% for testing. This method implemented using the 'train_test_split' function from scikit-learn with a random state of 42, provided a reliable foundation for evaluating the models' predictive accuracy. Figure 3 describes the model building flow chart for the work in general.

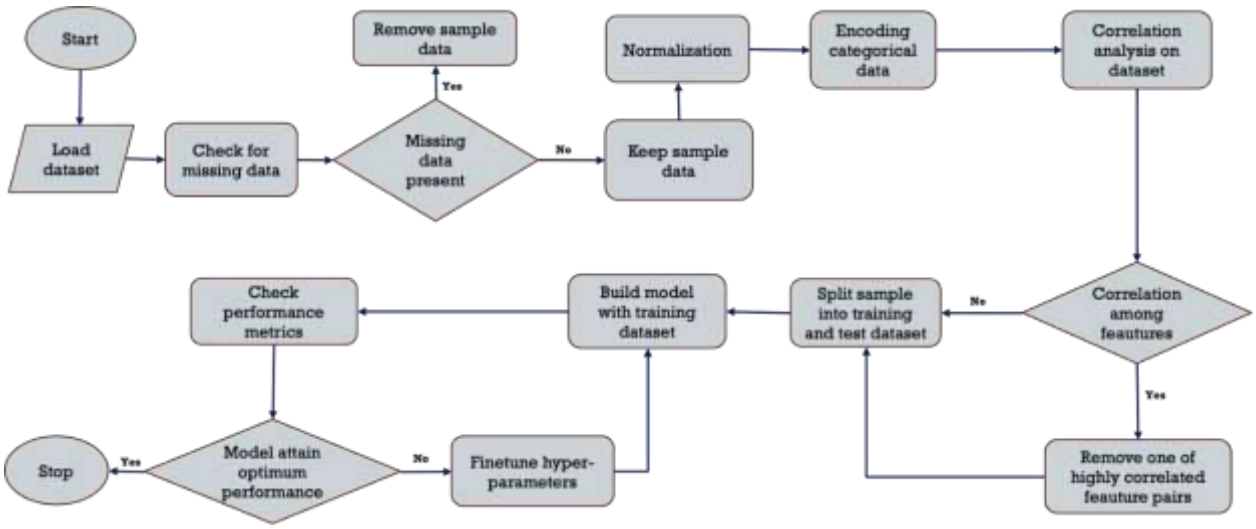



Figure 3: ML Model building flowchart

### 2.3 Model Training

In the development of our predictive models, we selected three powerful and diverse machine learning techniques. These are Support Vector Regression (SVR), Artificial Neural Networks (ANN) and Random Forest. Each of these models possesses different advantages in dealing with nonlinearity and large datasets, which are essential in predicting our reservoir flow capacity. SVR was selected because of its ability to capture complex nonlinear relationships in high dimensional spaces, ANN for its competence in learning intricate patterns, and Random Forest for its robustness in handling feature interaction and also preventing overfitting.

While other regression models such as Gradient Boosting Machines (GBM) and Extreme Gradient Boosting (XGBoost) are commonly used in prediction modelling, they were not included in this work due to computational considerations and focus on models with established interpretability in reservoir characterization [16]. For example, Random Forest provides a comparable advantage in capturing non linearity while being computationally more efficient on large datasets [17].
The same distribution of the train-test split consisting of 1372 and 343 data points respectively was used across all the algorithms.

#### 2.3.1 Support vector regression (SVR)

Support Vector Regression (SVR) is considered as a standard technique in supervised learning, particularly suited for linear fitting with data that may be non-linearly distributed. The SVR algorithm can perform the following functions:

1. **Linear Fitting:** SVR is fundamentally a method of linear fitting technique and works well only for linearly distributed data. In the case of nonlinear distributions, a specific nonlinear mapping is used on the data and the result is then used in a higher dimensional space where a linear method can be used to fit to the training data.
2. **Optimal Hyperplane Construction:** SVR aims at building the best hyperplane for the feature space through minimizing the regularized risk. Unlike traditional regression models, SVR incorporates a structural risk minimization framework that enhances its capability to achieve a global optimum, thereby reducing the risk of overfitting. Given a training sample set shown in Equation 1:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \text{ with } y_i \in R, \quad (1)$$

The objective of Equation 1 is to construct a regression model that closely predicts y shown in Equation 2.

$$f(x) = w^T x + b \quad (2)$$

Here, $w$ and $b$ are the model parameters that need to be optimized. Traditional regression minimizes the error between the predicted output $f(x)$ and actual value of $y$. In contrast, SVR introduces a margin of tolerance $\epsilon$, where no penalty is applied if the prediction error falls within this margin. The optimization problem can be formulated as:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} l_\epsilon(f(x_i) - y_i) \quad (3)$$

where C is a regularization constant, and l is the -insensitive loss function defined as:

$$l_\epsilon(z) = f(x) = \begin{cases} 0, & if\ |z| \leq \epsilon. \\ |z| - \epsilon, & otherwise. \end{cases} \quad (4)$$

In practical scenarios, kernel functions are employed to map the input data to a high-dimensional space, allowing the SVR algorithm to handle nonlinear relationships. Some commonly used kernels include:Linear kernel, Polynomial kernel, Gaussian kernel and Sigmoid kernel. These kernels facilitate the computation of inner products in higher-dimensional spaces, enabling SVR to model complex, nonlinear relationships effectively.

We varied different Kernel types, such as the Linear Kernel and Radial Basis Function (RBF) in assessing the distribution characteristics of our data-set. This was achieved by employing a systematic approach parameter tuning using GridSearchCV tool in Scikit-learn library. The GridSearchCV function is used for hyperparameter tuning in machine learning models. This involves setting the parameters for the kernel type, regularization parameter, epsilon etc. The parameter grid is defined as:

```
parameter_grid = {
        estimator_kernel:  ['linear','poly', 'rbf'],
       estimator_c:  np.range200, 500, 50 #Range from 200 to 500 with steps of 50,
      estimator_epsilon:np.range3, 7, 0.5 #Range from 3 to 7 with steps of 0.5
          }
      where estimator_kernel is the kernel type
      estimator_c is the regularization parameter
estimator_epsilon is the epsilon in the epsilon-SVR model
```

This approach enabled us to examine different ranges of the configurations in order to find the best combination. Before training the model, weights were adjusted optimally with the help of the parameter grid to reduce the loss function. Our SVR model was validated using an independent testing dataset by cross-validation. This validation step is critical since it helps in testing the model's ability to generalize in other unseen data.

### 2.3.2 Artificial neural network (ANN)

An Artificial Neural Network (ANN) emulates the data processing patterns observed in biological nervous systems like the human brain, albeit on a significantly smaller scale. The core concept involves devising novel information processing structures [18]. In this system, every connection has a specific weight, and each neuron is defined by a threshold value and an activation function [19]. The influence of an input's weight determines whether it has a positive or negative impact on the signal transmitted across a connection. A neuron will only relay a signal if the combined input exceeds its threshold. The activation value, a weighted sum, dictates the neuron's output, establishing a direct link between the weights of elements and the ANN's inputs and outputs, as illustrated in Figure 4.

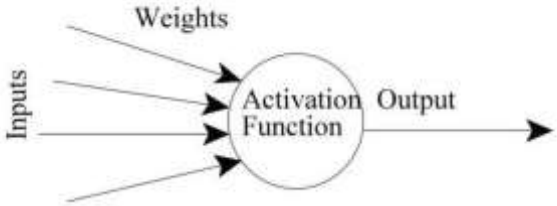



Figure 4: Inputs, weights of each input element and output of the ANN system.

ANNs can be categorized into two topologies: Feed Forward and Feedback. Feed Forward systems lack feedback loops; information only travels in one direction, ensuring that inputs and outputs remain static. Each unit takes an input from the previous units subsequently and weights the inputs by the connections to decide the output outcome based upon the weights it has received. On the other hand, Feedback ANNs use content addressable memories in learning process steps. This involves the changing of the weights between the connection weights based on the difference between the output of the network and the expected output [20]. With this feedback mechanism, weights may be adjusted dynamically to enhance the network's accuracy and also its performance.

We trained the model through backpropagation algorithm wherein the weights of the network were adjusted to reduce the error difference between the predictions from the actual output. This process was implemented using the TensorFlow's

Keras architecture. The model was initiated as a sequential stack of neural network layers. We employed the ReLU activation function. This was chosen because of its effectiveness in adding non-linearity to the model and also reducing vanishing gradients. The input layer constituted 128 neurons, the hidden layers were varied to derive the optimal number for our use case, and the output layer was a dense linear layer typical for regression tasks. In order to achieve stability and efficiency in the training process, we used the 'adam' optimizer to a set clipnorm of 1. 0 for the purpose of gradient clipping. Then, we integrated the ReduceLROnPlateau callback as a dynamic learning rate scheduler for the training process. This scheduler automatically reduces the learning rate by a factor of 0.2 whenever the validation loss shows no improvement for 10 consecutive epochs. The training process also included an early stopping mechanism set with patience of 30 epochs. This callback was added to stop the training process when the validation loss was no longer improving and also retrieving the best model weights during the training.

### 2.3.3 Random forest

These models are based off tree-like model of decision and its consequence, which makes them intuitive and easy to interpret. The decision trees developed as a way of constructing discriminative models are one of the oldest and the most popular methods applied in both statistical and machine learning fields, evolving independently across the disciplines [21]. The Bagging algorithm developed by Leo Breiman set the foundational groundwork for the inception of Random Forests. Bagging (or Bootstrap Aggregating) is a process of generating a predictor-like decision tree in multiple versions on various subsets of original data, which are then sampled with replacements and combines the versions together to form better prediction accuracy. Based on this concept, the Random Forests include another layer of randomness in the process of creating trees alongside with the use of bootstrapped samples. Random Forests (RF) is a rather unique ensemble machine learning algorithm which encompasses a series of tree classifiers. In this ensemble, each tree votes towards identifying the most prevalent class, the votes are then aggregated together in order to produce the prediction outcome [22]. Thus, Random Forests are characterized by high accuracy. They are less sensitive to noise, outliers and usually do not over fit data.

We used the Random Forest Regressor class from the scikit-learn package and tuned hyperparameters that control the complexity of the model. These are: number of trees present in the forest (*n_estimators*), the maximum depth of each tree (*max_depth*) and minimum number of samples required to split an internal node (*min_samples_split*).
A grid search technique was employed to explore a range of values for these hyper-parameters. In this study, we set ranges of our grid parameters for the *n_estimators* and *min_samples_split* to [50, 100, 200, 300, 400, 500] and [2, 5, 10, 20] respectively. The optimal hyperparameters obtained from the grid search were then used to train the model.

### 2.4 Model Evaluation

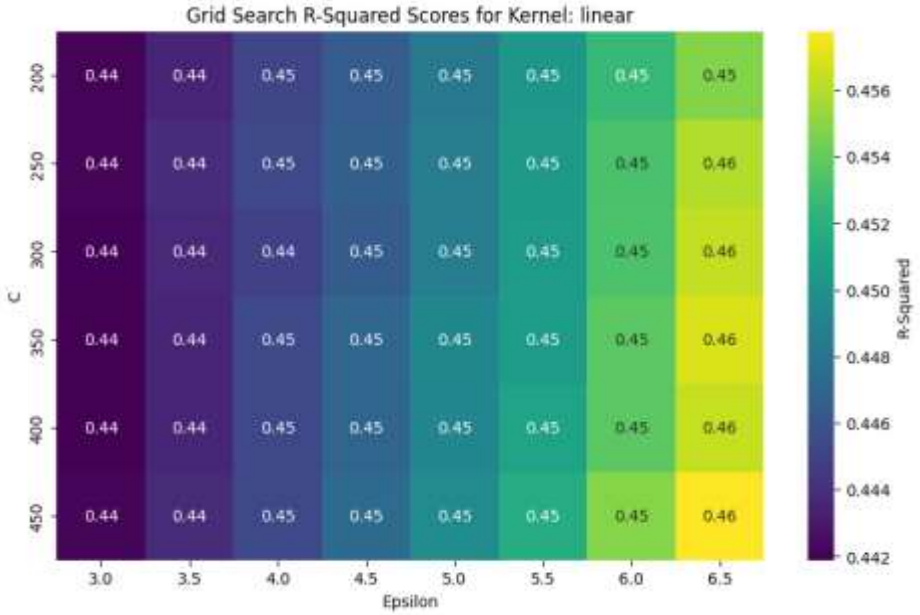
The models were evaluated based on three performance metrics, namely: R-Squared, Mean Absolute Error (MAE), Mean Square Error (MSE), and Explained variance. These metrics tell us the overall fit of the model, provide an average error magnitude and focus on the portion of variability the model accounts for respectively.

1. **R-Squared** $(R^2)$: also known as the coefficient of determination, provides an indication of how well the model explains the observed data, it measures the proportion of the variance in the dependent variable that is predictable from the independent variables with a value of 1 meaning the model explains all variability and 0 meaning it explains none. The formula is shown in Equation 5.
   a. $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$
   b. where: $y_i$ is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, $n$ is the number of observations.
2. **Mean Absolute Error (MAE)**: provides a straightforward measure of prediction accuracy by showing how close predictions are to the actual values on average, i.e. the average of the absolute errors between predicted and actual values with lower values indicating better performance. The formula for calculating MAE is shown in Equation 6.
   a. $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$
   b. Where: $y_i$ is the actual value, $\hat{y}_i$ is the predicted value, $n$ is the number of observations.
3. **Mean Squared Error (MSE):** as shown in Equation 7 measures the cumulative average squared difference between the actual and predicted values. Unlike MAE, MSE penalizes larger errors more heavily because of the squaring, making it particularly sensitive to outliers. A lower MSE indicates better model performance.
   a. $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$
   b. Where: $y_i$ is the actual value for observation $i$, $\hat{y}_i$ is the predicted value for observation $i$, $n$ is the number of observations.
4. **Explained Variance**: quite similar to $R^2$, measures the proportion of the variance in the target variable that the model explains. It however focuses solely on how well the model captures variability, without penalizing for the model's prediction bias. The formula is shown in Equation 8.
   $$Explained\ Variance = 1 - \frac{Var(y - \hat{y})}{Var(y)} \quad (8)$$
   where: $Var(y)$ is the variance of the actual values, $Var(y - \hat{y})$ is the variance of the errors (actual - predicted values).

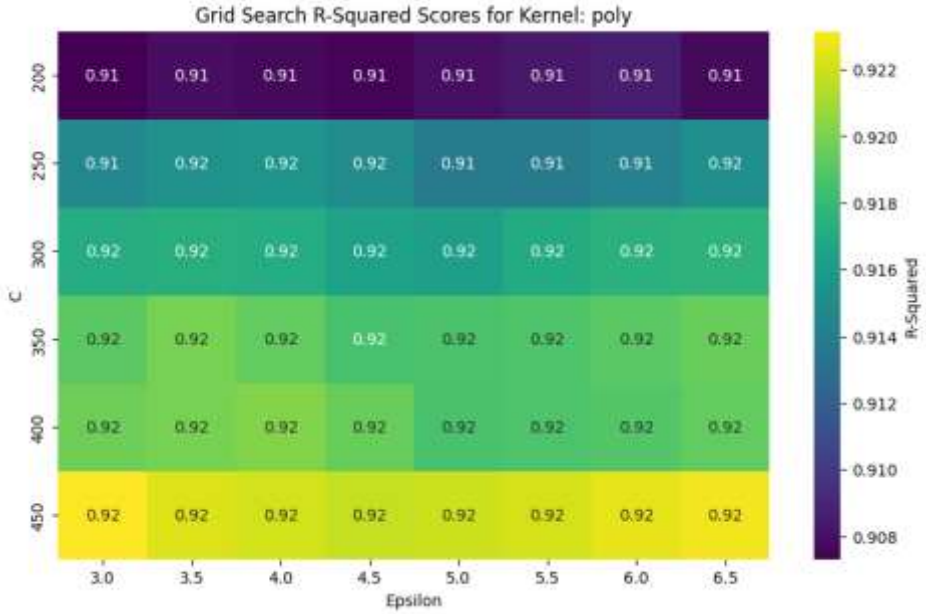
## 3. RESULTS AND DISCUSSION

Each of the models developed was tested to predict Reservoir Flow Capacity (RFC) based on a selection of geophysical attributes. We present the evaluation of each model's performance by analyzing their predictive accuracy, the relevance of the features they considered and their ability to handle the complexity of the features. Feature importance analysis was conducted for the RF and SVR models due to their inherent ability to rank input variables. However, the ANN model does not readily provide feature importance because it operates a black box model; hence, this analysis was not performed.

**SVR:** Upon evaluation of our grid search, as shown in Figure 5a, 5b and 5c, we found the optimal kernel type, regularization and epsilon parameter, as: (kernel='poly', C=450, epsilon=3.0) and our SVR model was trained on this.

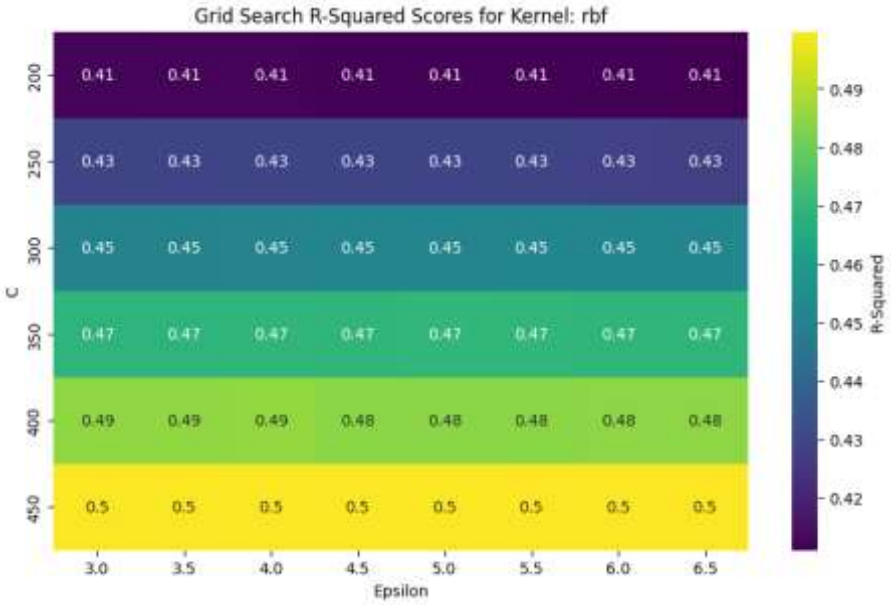



(a): Grid search results for the linear kernel, showing relatively low R-squared values across all hyperparameter combinations.




(b): Grid search results for the polynomial kernel, which demonstrates significantly higher R-squared values compared to the linear and RBF kernels.

Figure 5: R-Squared scores for SVR Hyper-Parameter Grid Search




(c): Grid search results for the RBF kernel, showing moderate performance improvement with increasing C values.

Figure 5: R-Squared scores for SVR Hyper-Parameter Grid Search (Cont'd)

An insightful feature importance analysis revealed in Figure 6 that geological formation thickness and reservoir thickness emerged as highly influential features, garnering importance levels of 8 and 9, respectively, on a scale of 1 to 10. This shows the model's inclination towards prioritizing these features in its predictions. Conversely, features such as gas permeability and liquid permeability, with importance levels of 0.1 each, were deemed less impactful in the model's decision-making process.

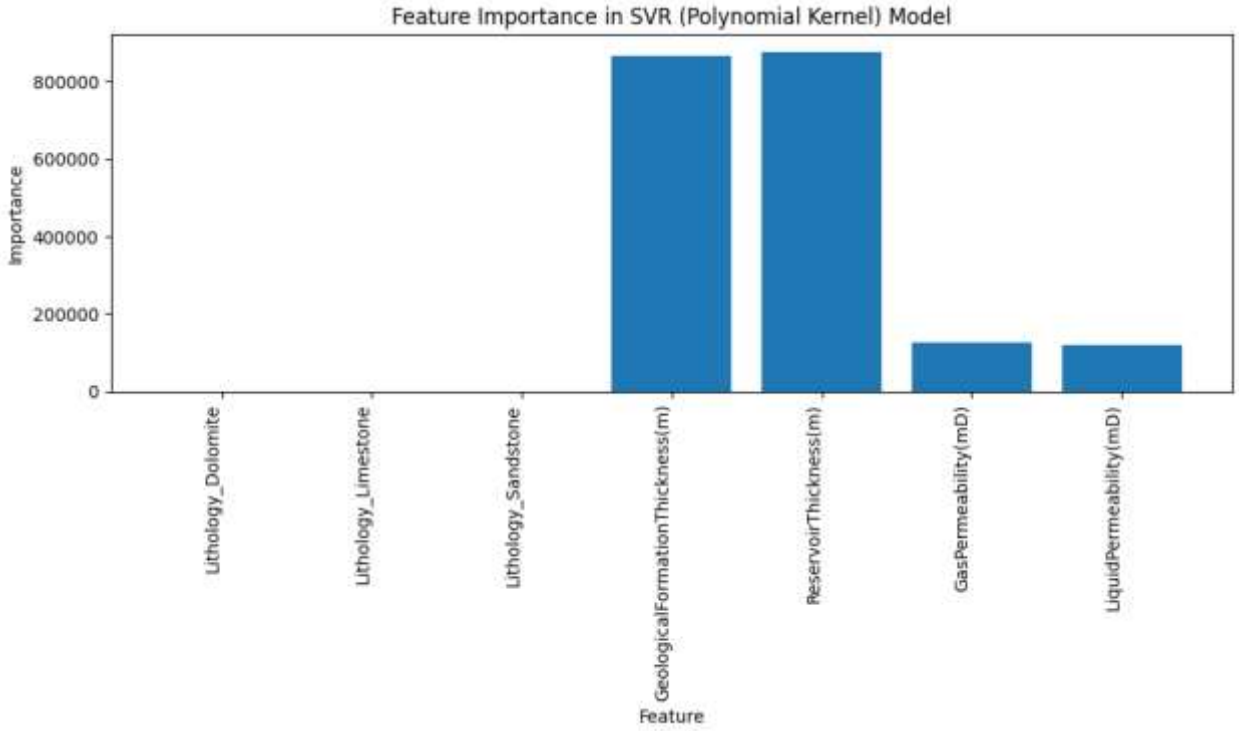



Figure 6: SVR feature importance of geological features to target variables

**ANN:** The variation in the number of hidden layers in the ANN model was centered on optimizing its training process. The training and validation loss reduction was recorded over epochs, and its ability to model intricate patterns not readily apparent to other models. Each ANN model training was varied with increasing numbers of hidden layers in steps of one,

i.e. we trained each model with 1-10 hidden layers respectively. Each model had varying epochs as the number of hidden layers varied because of the early stopping mechanism we implemented in the methodology. Each ANN model trained on its varying epochs and number of hidden layers also predicted different outputs and yielded different results. The optimum network configuration had 9 hidden layers with an R-Squared value of 0.939 shown in Figure 7 and Mean Square Error of 1065 shown in Figure 8.

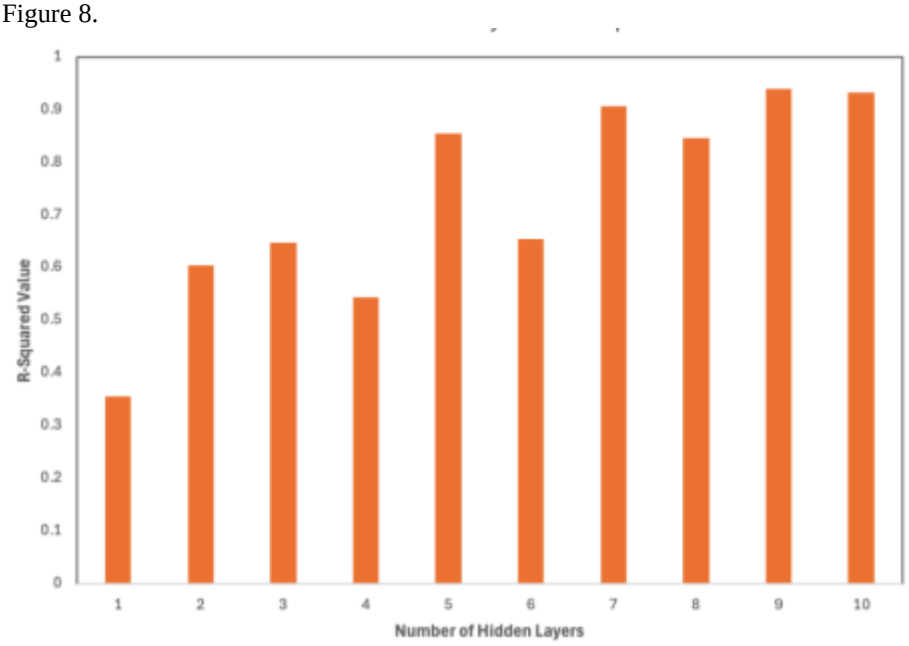



Figure 7: Relationship between the number of ANN hidden layers and the R-Squared value

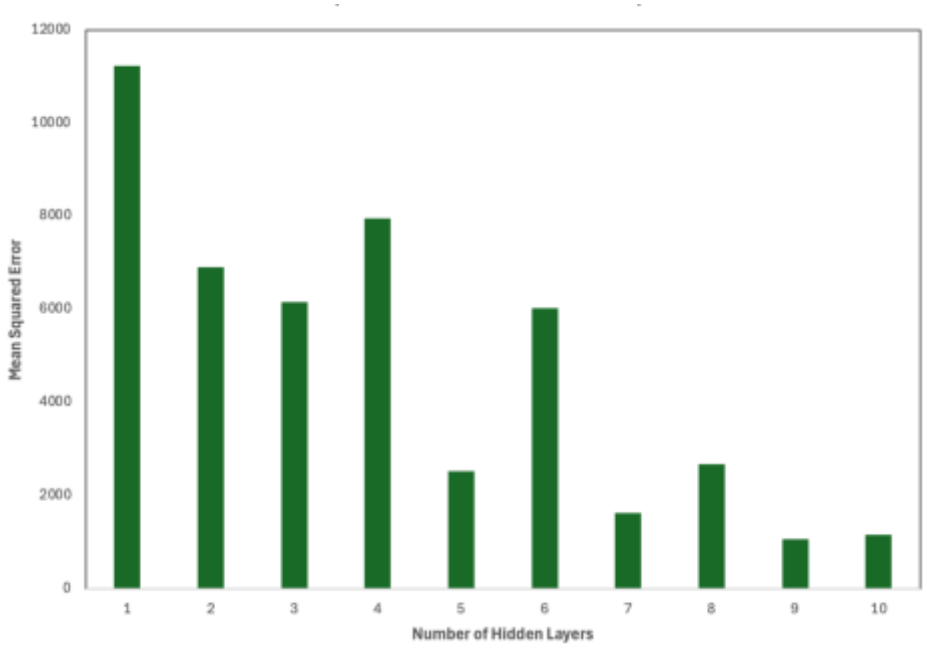



Figure 8: Relationship between the number of ANN hidden layers and the MSE value

**RF:** We also examine the results from the grid search for our Random Forest model optimal hyper-parameters. The outcome of the grid search showed that the optimal number of estimators and minimum samples split for our data are 100 and 2 respectively as shown in Figure 9.

The feature importance plot for the Random Forest model in Figure 10 provides further insights into which features are most influential in the prediction of RFC values. It shows that geological formation thickness and liquid permeability are the most significant predictors, holding the highest importance scores. The importance of liquid permeability reiterates the fact that the ease with which a fluid passes through a reservoir's pores is intuitively a determining factor for the flow capacity. Likewise, the thickness of the geological formation points to the volume available for fluid storage and flow, reinforcing its relevance.

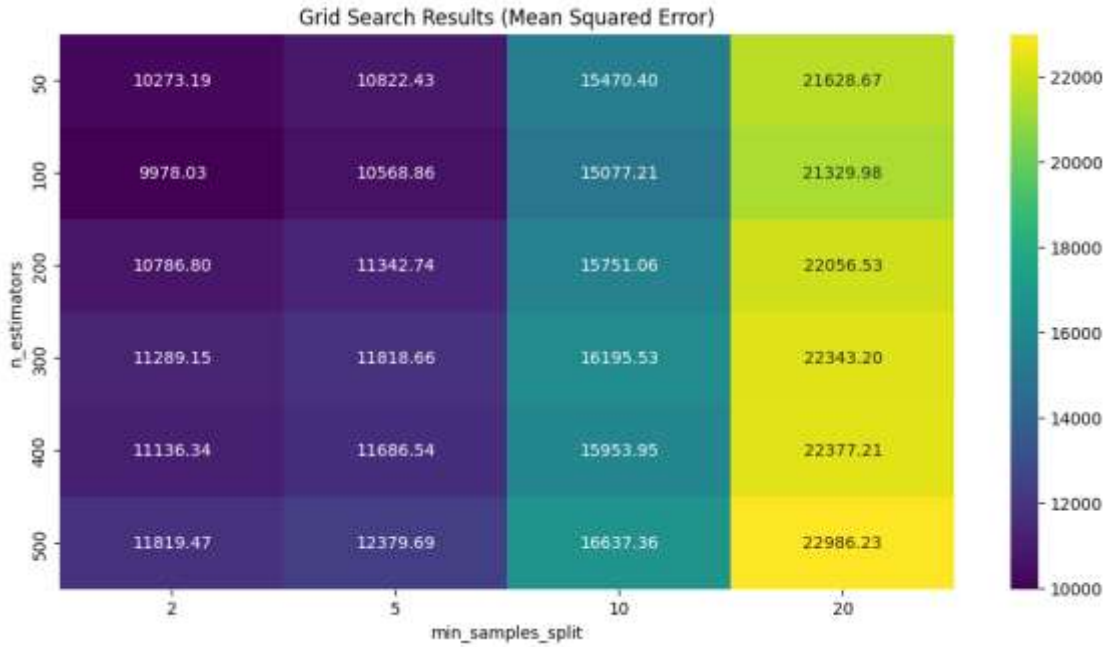



Figure 9: MSE scores for the random forest hyper-parameters grid search

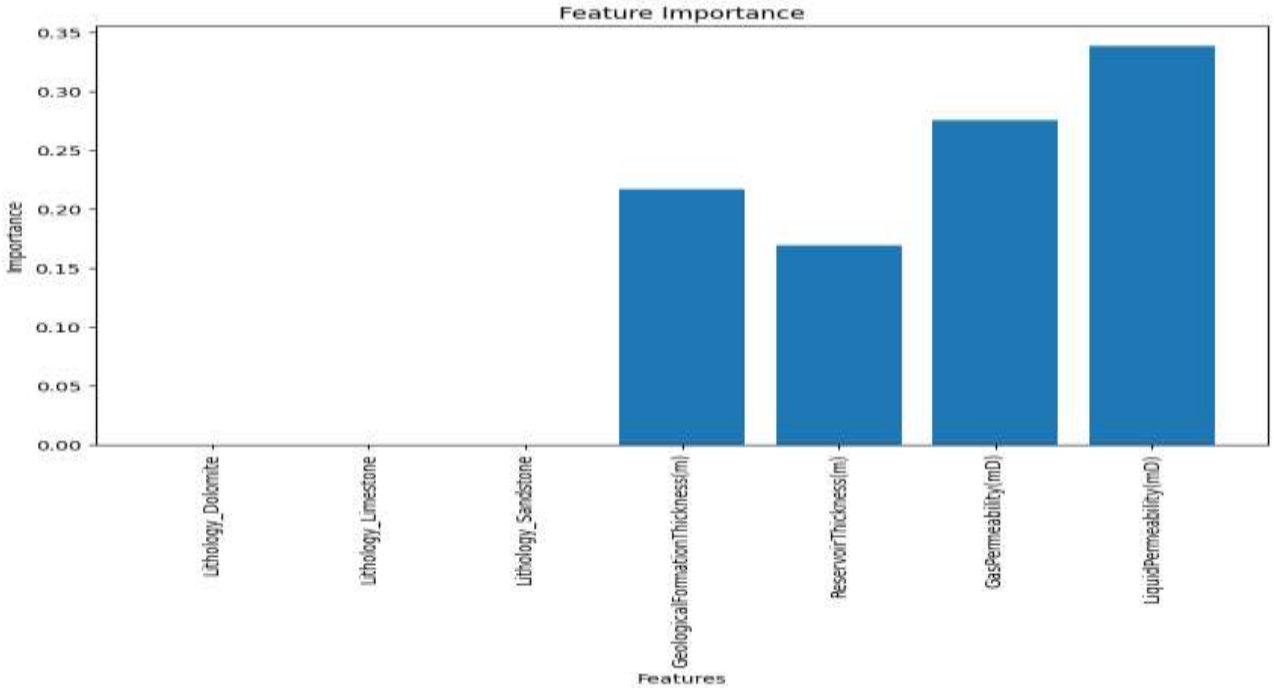



Figure 10*:* Random forest feature importance of geological features to target variables.

Table 3 summarizes the final selected hyperparameters from grid search for SVR and Random Forest while it also contains the number of hidden layers used for the ANN model.

Table 3: Summary of the optimal hyper-parameters

| ML Algorithm | Kernel | C (Regularization Parameter) | Epsilon | No of estimators | Minimum samples split | No of hidden layers |
|---|---|---|---|---|---|---|
| SVR | poly | 450 | 3.0 | N/A | N/A | N/A |
| ANN | N/A | N/A | N/A | N/A | N/A | 9 |
| RF | N/A | N/A | N/A | 100 | 2 | N/A |

The feature importance analysis carried out using the SVR (Figure 7) and RF (Figure 11) models revealed that geological formation thickness, reservoir thickness and permeability (gas and liquid) are the most critical factors in predicting RFC. These findings underscore the importance of geological and petrophysical characteristics in reservoir performance and mechanical stability. The dominance of geological formation thickness indicates its strong influence on the structural integrity and fluid migration pathways within the reservoir. Reservoir thickness which was also found to be important determines the volume of hydrocarbon-bearing rock and influences pressure dynamics during production. Additionally, gas and liquid permeability exhibited high importance in both models, highlighting the direct impact on fluid mobility, production rates and overall reservoir efficiency.

These observations align with recent studies emphasizing the importance of geological and petrophysical properties in reservoir characterization. For instance, Osahon et al., [23] demonstrated that integrating geological formation thickness and permeability data enhances the accuracy of reservoir models. Similarly, Solanke et al [24] highlighted that advanced geological modeling techniques, which incorporate factors like formation thickness and permeability, significantly improve predictions of reservoir performance.

By assessing the three models based on key performance metrics – R-Squared, Mean Absolute Error, and Explained Variance, the study found that both the Random Forest and Artificial Neural Network (ANN) models exhibited strong capability in capturing the variability of reservoir flow capacity, as evidenced by their high R-squared values. In contrast, the SVR model demonstrated lower performance, as shown in Table 4. These results indicate that the Random Forest model achieved the best fit to the dataset. The lower performance of the SVR model showed that the algorithm poorly captured the complex nonlinear relationships present in the dataset. This could be because of the sensitivity of SVR to hyperparameter selection and distribution of data points which may not have effectively mapped the geological and petrophysical features influencing RFC. The superior performance of RF as seen in Table 4 with the highest R-squared value (0.9573) and lowest Mean Absolute Error (4.6545) can be attributed to its ensemble learning approach which combines multiple decision trees to improve predictive accuracy and reduce overfitting. From Figure 11 which shows the plot of the errors for the different models, it is evident that the random forest model was able to capture the patterns and relationships in the dataset better than ANN and SVR.

Table 4: Performance metrics of the RFC prediction models

| ML Algorithm | R-Squared | Mean Absolute Error | Mean Squared Error | Explained Variance |
|---|---|---|---|---|
| SVR | 0.7294 | 15.2871 | 4727.52 | 0.7319 |
| ANN | 0.9390 | 9.0048 | 1065.31 | 0.9412 |
| Random Forest | 0.9573 | 4.6545 | 746.13 | 0.9578 |

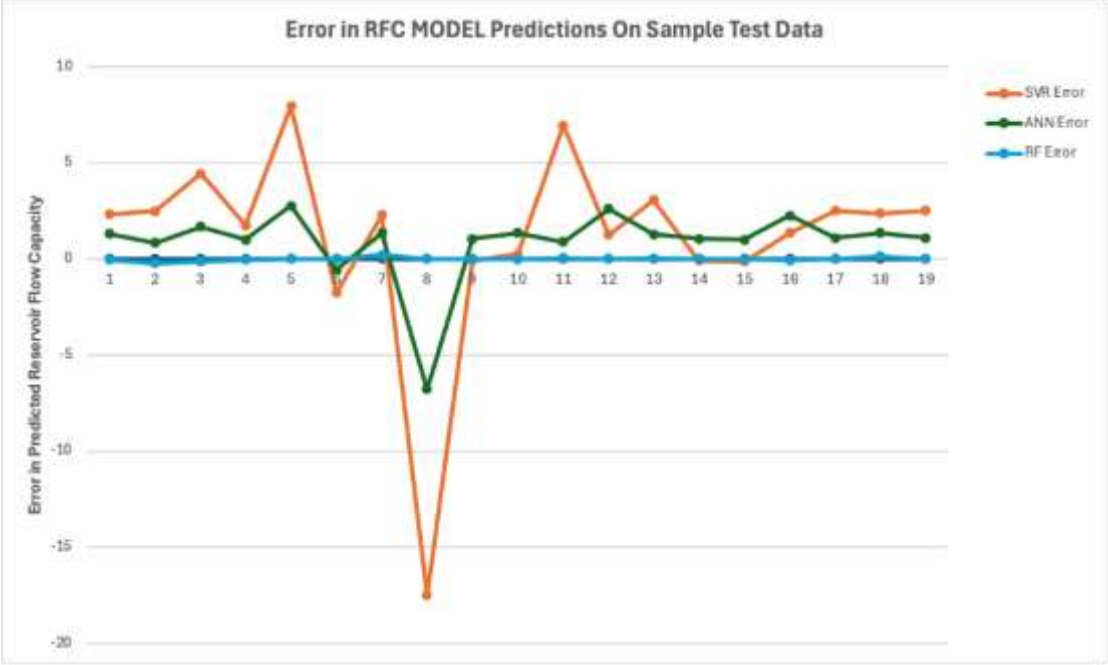



Figure 11: Prediction errors of the RFC model on sample test data

## 4. CONCLUSION AND FUTURE WORKS

Understanding sand occurrence requires a multidisciplinary approach, by combining knowledge of geology, petroleum engineering, and fluid dynamics, among others. This research highlights the potential of machine learning models in predicting reservoir flow capacity and assessing factors contributing to sand production. The results validate the

effectiveness of machine learning in capturing complex relationships between geological properties and reservoir behavior, reinforcing the importance of thickness and permeability as critical predictors of reservoir flow capacity and stability.
Higher accuracy values recorded in the RF and ANN models can be attributed to the abilities of the models in handling complex, non-linear relationships between features. Random forest uses multiple decision trees aggregation to capture complex patterns in the data and ANNs use multilayer architecture and activation functions to unravel these patterns. On the contrary, the SVR models are predominantly linear models, even though applying polynomial kernels introduces non-linearity, it does not effectively capture the non-linearity as observed in RF or ANN.

In future work, the primary focus will be on acquiring extensive real-time reservoir data from oil and gas industries. This will be important in improving the ability of the models to predict accurately and aid robustness of the models. The feasibility of real-time data acquisition will be carefully evaluated, considering industry data availability, sensor technology and integration with existing monitoring systems. Through incorporating real time data, the models can be updated and refined continuously, making it possible to get better predictions of the reservoir flow capacity and enhanced control of sand production over time during the drilling processes. Furthermore, there's a need to expand the scope of the variables we considered in our model. Introducing additional geological and petrophysical parameters that contribute to reservoir flow capacity will provide a better understanding of factors influencing the generation of sand and its prediction thereby leading to enhanced and targeted sand management strategies in drilling operations.

**Abbreviations:**
ANN: Artificial Neural Network
GDR: Geothermal Data Repository
ML: Machine Learning
MAE: Mean Absolute Error
MSE: Mean Square Error
RFC: Reservoir Flow Capacity
RF: Random Forest
RFC: Random Forest Classifier
RBF: Radial Basis Function
ReLU: Rectified Linear Unit
SVM: Support Vector Machine
SVR: Support Vector Regression

## REFERENCES

1. Halsey, T., Agrawal, G., Bailey, J. R., Balhoff, M., Borglum, S. J., Mohanty, K. K., & Traver, M., (2023). Grand challenges for the oil and gas industry for the next decade and beyond. *Journal of Petroleum Technology*.
2. Ayal, A., & Sadeq, D. (2024). Sand production prediction and management using a one-dimensional geomechanical model: A case study in NahrUmr Formation, Subba Oil Field. *Iraqi Geological Journal, 57*(2C), 64–83. https://doi.org/10.46717/igj.57.2C.6ms-2024-9-14
3. Laoufi, H., Megherbi, Z., Zeraibi, N., Merzoug, A., & Ladmia, A. (2022). Selection of sand control completion techniques using machine learning. *International Geomechanics Symposium*. https://doi.org/10.56952/IGS-2022-118
4. Chen, X., Zhang, J., Wang, Y., Liu, H., Li, Q., & Wu, Y. (2024). A real-time drilling parameters optimization method for offshore large-scale cluster extended reach drilling based on intelligent optimization algorithm and machine learning. *Ocean Engineering, 291*, 116375. https://doi.org/10.1016/j.oceaneng.2023.116375
5. Arab Oil & Natural Gas. (2017). *Sand control - AONG*. https://www.arab-oil-naturalgas.com/sand-control/
6. Saghandali, F., Salehi, M., Hosseinzadehsemnani, R., Moghanloo, R. G., & Taghikhani, V. (2022). A review on chemical sand production control techniques in oil reservoirs. *Energy & Fuels, 36(10), 5185-5208*. https://doi.org/10.1021/acs.energyfuels.2c00700
7. Kazidenov, D., Omirbekov, S., Zhanabayeva, M., & Amanbek, Y. (2025). Experimental and numerical study of the effect of polymer flooding on sand production in poorly consolidated porous media. *Geoenergy Science and Engineering, 249*, 213746. https://doi.org/10.1016/j.geoen.2025.213746
8. Muñoz-Ibáñez, A., Herbón-Penabad, M., Li, Y., & Delgado-Martín, J. (2025). Impact of fluids on the Mode I fracture toughness of two granites and one sandstone. *Journal of Geophysical Research: Solid Earth, 130*(1), e2024JB030441. https://doi.org/10.1029/2024JB030441
9. Asfha, D. T., Abubakar, A. I., Ahmed, M. A., Shariff, A. M., & Kamel, A. R. (2024). Mechanisms of sand production, prediction–a review and the potential for fiber optic technology and machine learning in monitoring. *Journal of Petroleum Exploration and Production Technology, 14*(10), 2577–2616. https://doi.org/10.1007/s13202-024-01860-1
10. Cheddad, F. A. (2023). Enhancing petrophysical studies with machine learning: A field case study on permeability prediction in heterogeneous reservoirs. *arXiv*. https://doi.org/10.48550/ARXIV.2305.07145
11. Krishna, S., Irfan, S. A., Keshavarz, S., Thonhauser, G., & Ilyas, S. U. (2024). Smart predictions of petrophysical formation pore pressure via robust data-driven intelligent models. *Multiscale and Multidisciplinary Modeling, Experiments and Design, 7*(6), 5611–5630. https://doi.org/10.1007/s41939-024-00542-z

12. Ali, M., Zhu, P., Jiang, R., Huolin, M., Ehsan, M., Hussain, W., Zhang, H., Ashraf, U. & Ullah, J. (2023). Reservoir characterization through comprehensive modeling of elastic logs prediction in heterogeneous rocks using unsupervised clustering and class-based ensemble machine learning. *Applied Soft Computing, 148*, 110843. https://doi.org/10.1016/j.asoc.2023.110843
13. Otmane, M., Imtiaz, S., Jaluta, A. M., & Aborig, A. (2025). Boosting reservoir prediction accuracy: A hybrid methodology combining traditional reservoir simulation and modern machine learning approaches. *Energies, 18*(3), 657. https://doi.org/10.3390/en18030657
14. Jordan, T. E. (2016). Appalachian Basin play fairway analysis: Natural sedimentary reservoirs data 2016 revision [data set]. *DOE Geothermal Data Repository; Cornell University*. https://doi.org/10.15121/1495428
15. Raheem, E. (2024). Missing data imputation: A practical guide. In A. K. Mitra (Ed.), *Statistical Approaches for Epidemiology: From Concept to Application* (293–316). Springer International Publishing. https://doi.org/10.1007/978-3-031-41784-9_18
16. Min, C., Wen, G., Gou, L., Li, X. & Yang, Z. (2022). Interpretability and causal discovery of the machine learning models to predict the production of CBM wells after hydraulic fracturing. *arXiv*. https://doi.org/10.48550/ARXIV.2212.10718
17. Feng, P., Wang, R., Sun, J., Yan, W., Chi, P., & Luo, X. (2024). An interpretable ensemble machine-learning workflow for permeability predictions in tight sandstone reservoirs using logging data. *Geophysics, 89*(5), MR265–MR280. https://doi.org/10.1190/geo2023-0657.1
18. Onciul, R., Tataru, C.-I., Dumitru, A. V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Radoi, M. P., & Toader, C. (2025). Artificial intelligence and neuroscience: Transformative synergies in brain research and clinical applications. *Journal of Clinical Medicine, 14*(2), 550. https://doi.org/10.3390/jcm14020550
19. Putra, D. S., Azmi, M., Muslikhin, & Purwanto, W. (2022). ANN activation function comparative study for sinusoidal data. *Journal of Physics: Conference Series, 2406*(1), 012029. https://doi.org/10.1088/1742-6596/2406/1/012029
20. Saha, N., Swetapadma, A., & Mondal, M. (2023). A brief review on artificial neural network: Network structures and applications. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (1974–1979). IEEE. https://doi.org/10.1109/ICACCS57279.2023.10112753
21. Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylon Journal of Machine Learning, 2024*, 69–79. https://doi.org/10.58496/BJML/2024/007
22. Becker, T., Rousseau, A.-J., Geubbelmans, M., Burzykowski, T., & Valkenborg, D. (2023). Decision trees and random forests. *American Journal of Orthodontics and Dentofacial Orthopedics, 164*(6), 894–897. https://doi.org/10.1016/j.ajodo.2023.09.011
23. Osahon, U., & Efetobore, G. M. (2023). Reservoir characterization: Enhancing accuracy through advanced rock physics techniques. *Journal of Geosciences and Geomatics, 11*(2), 67–78. https://doi.org/10.12691/jgg-11-2-4
24. Solanke, B., Onita, F. M., Ochulor, O. J., & Iriogbe, H. O. (2024). Techniques for improved reservoir characterization using advanced geological modeling in the oil and gas industry. *International Journal of Applied Research in Social Sciences, 6*(9), 2060–2088. https://doi.org/10.51594/ijarss.v6i9.1542