**ANL 488 PROJECT PROPOSAL**

**Using Artificial Neural Network to predict Oil and Gas**

**Recovery Factor**



**Submitted by**

**Ho Kai Toong**

**SCHOOL OF BUSINESS Singapore University of Social Sciences**

**Presented to Singapore University of Social Sciences in partial fulfillment of the requirements for the**

**Degree of Bachelor of Science in Business Analytics**

**2023**

Table of Contents

[Introduction 1](#_Toc144133115)

[Literature Review 3](#_Toc144133116)

[Data Understanding and Preparation 6](#_Toc144133117)

[TORIS 6](#_Toc144133118)

[GOM 9](#_Toc144133119)

[Proposed Modelling and Evaluation 11](#_Toc144133120)

[Proposed Schedule 12](#_Toc144133121)

[References 13](#_Toc144133122)

# Introduction

The prediction of recovery factor (RF) is a critical process in the oil and gas (O&G) sector. RF is typically measured by the proportion of O&G available for extraction over the total volume of hydrocarbons in a reservoir. It is one of the most significant parameters for O&G development as it directly reflects how much of the hydrocarbons can be economically recovered in a reservoir. In other words, the higher the RF, the higher the efficiency of oil production from the reservoir (Afari et al., 2015).

An accurate estimation of RF can also help O&G businesses to discover factors that lead to a higher rate of RF which can generate a higher return on investments. Therefore, by accurately predicting the RF, O&G companies can effectively strategize their resources to optimize oil production and revenue (Sachin & Karmakar, 2020).

However, due to the composite nature of oil reservoirs, predicting O&G RF can be challenging. The heterogeneity of oil reservoirs, reservoir pressure, and rock parameters such as porosity and permeability tend to create unpredictable trends on the O&G recovery (Ahmed et al., 2019). In the early days, the O&G industry encountered more difficulties in achieving accurate predictions of RF due to the lack of production data. As a result, operators in the O&G sector had to rely on their field experiences and various analogue studies to predict the RF (Kumar et al., 2022). Thus, an efficient predictive technique is essential to tackle the complexity of estimating O&G RF.

The artificial neural network (ANN) model is suitable in predicting O&G RF. ANN is known to discover complex, nonlinear, and multi-dimensional functional relationships between the inputs and output. Given the intricate nature of oil reservoirs, ANN is immensely appropriate for the prediction of O&G RF. Furthermore, predicting O&G RF usually requires regression predictive modelling which is applicable to ANN as it can predict a continuous output (Chan et al., 2022).

Moreover, studies have shown an increasing usage of AI to predict O&G RF (Li et al., 2020). Using ANN to predict O&G RF has also produced good predictive results which will be discussed in the literature review section.

With these, the objective of this project is to adopt ANN to predict O&G RF and obtain the optimized result by comparing and evaluating a series of ANN models.

# Literature Review

A study adopted the use of predictive analytics to predict the ultimate recovery factor (URF) of oil reservoirs in Gulf of Mexico. The dataset consisted of 4000 oil reservoirs with 82 attributes for each reservoir (Gowtham & Wu, 2017).

The predictive techniques adopted in the study consisted of multiple linear regression, robust linear regression, least absolute shrinkage and selection operator (LASSO), K-nearest neighbors, decision tree, random forest, ANN, and ensemble model. Root mean square error (RMSE) and mean absolute error (MAE) were used as evaluation methods on the testing set to assess the performance of the models.

Based on the modelling results, excluding the ensemble model, ANN emerged to be the champion model with the lowest RMSE and MAE of 8.5% and 6.0% respectively. Multiple linear regression is ranked as the second best model with RMSE and MAE of 9.0% and 7.0% respectively. And robust linear regression is the third best model with RMSE and MAE of 9.3% and 7.2% respectively. Decision Tree is the worst performing model with RMSE and MAE of 13.2% and 10.6% respectively.

It is observed that the champion model contained lesser errors compared to the other predictive models. The construction of the ANN model is optimized at 1 hidden layer and 3 nodes in it. The results also show that 80% of the predictions on the testing set contain an error of no more than 8%.

A further study adopted the use of artificial intelligence (AI) to predict the oil recovery factor (ORF) for Water Derive Sandy Reservoirs. The dataset involved 130 water drive sandstone reservoirs and an additional 38 reservoirs of data were collected to test the predictive performance of the models (Ahmed et al., 2019).

Four AI models of ANN, radial basis neuron networks (RNN), adaptive neuro-fuzzy inference system with subtractive clustering (ANFIS-SC), and support vector machines (SVM) were used to predict the ORF for this study. The evaluation measures were coefficient of determinations (R²) and absolute average percentage error (AAPE).

Among the four AI models, the ANN model has the best performance on the testing data with the highest R² of 0.94 and lowest AAPE of 7.92%. ANFIS-SC is the second best performing model with R² of 0.91 and AAPE of 8.53%, followed by RNN with R² of 0.88 and AAPE of 8.78%, and then SVM with R² of 0.90 and AAPE of 10.44%.

The construction of the best performing ANN model includes one hidden layers with 5 neurons in it. It is also observed that the trainlm (Levenberg–Marquardt) function is adopted during the training process and tan-sigmoid is used as the activation function in the output layer. Subsequently, the trained model is applied to establish the empirical correlation and predict the ORF and hence achieving the best R² and AAPE results.

Another study utilized AI to predict the movable ORF of a layered reservoir as a result of waterflooding. A three-dimensional simulation model was built and contained 10 layers and a total of 64,000 cells (Shams et al., 2020).

Three AI models were used to perform the prediction of ORF which included ANN, Non-linear Regression (NLR), and Adaptive neuro-fuzzy inference system (ANFIS). The mean absolute percentage error (MAPE) and R² were used to evaluate and compare the predictive models’ performance.

Based on the models’ performance comparison, the ANN model is identified as the best model in performance which achieved MAPE of 7.08% and R² 0.997 on the testing set. The second best model is ANFIS which obtained MAPE and R² of 9.47% and 0.996 respectively. The worst performing model is NLR with MAPE and R² of 34.23% and 0.950 respectively.

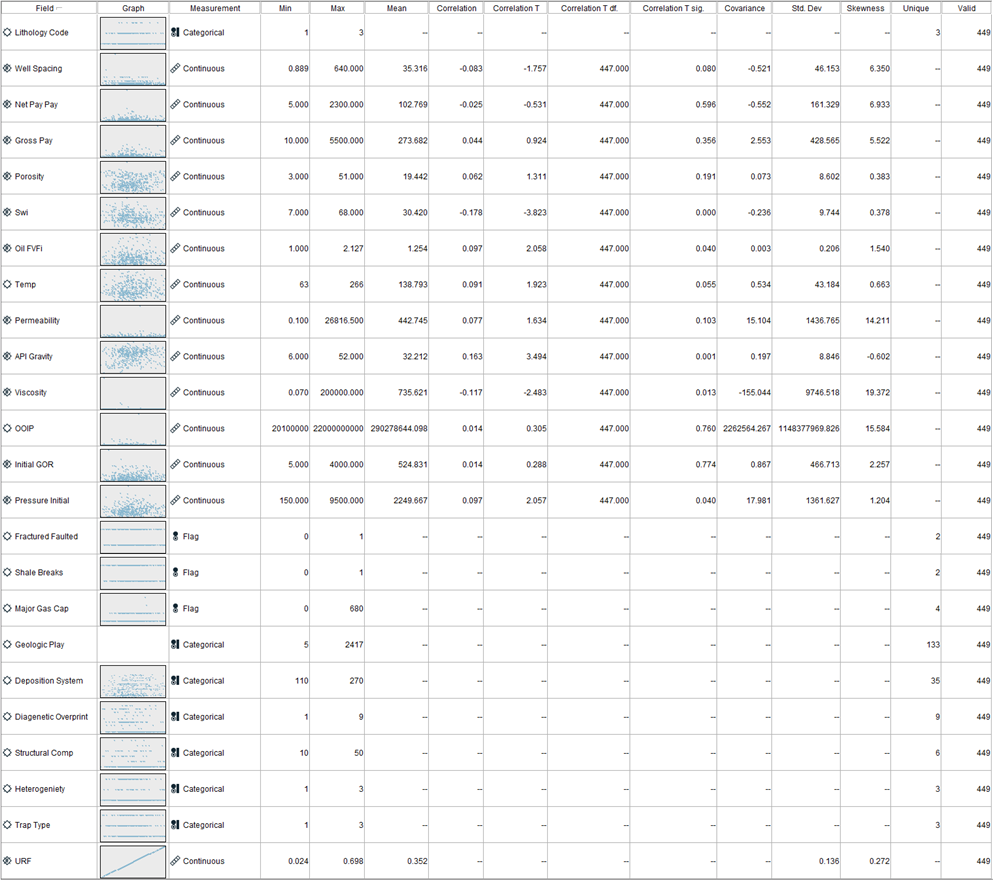
The construction of the ANN model is optimized at a hidden layer with 14 neurons. Similar to the previous study, the Levenberg–Marquardt function is used to train the model and tan-sigmoid is applied as the activation function in the output layer.

In addition, two further studies have adopted the use of ANN to predict O&G RF and produced decent predictive results (Afari et al., 2015; Surajudeen et al., 2022). Also, another study concluded that ANN is one of the most popular predictive techniques to predict the O&G R&F (Sachin & Karmakar, 2020). Therefore, this project aims to leverage the capabilities of ANN to predict the O&G RF.

# Data Understanding and Preparation

There are two datasets used for this study. The Tertiary Oil Recovery System (TORIS) dataset (U.S Department of Energy, 1995) which contains 449 records with 24 variables and the Gulf of Mexico (GOM) dataset (Bureau of Ocean Energy Management, 2019) which consists of 4,512 records with 15 variables.

## TORIS

****

**Figure 1. Statistics Table of TORIS Variables**

Referring to figure 1, there are 10 categorical and 13 continuous inputs used to predict a continuous output “URF”. It is also observed that the continuous variables contain values of different ranges. For instance, “Temp” ranges from 63 to 266 and “Viscosity” ranges from 0.07 to 200,000. To prevent this from affecting the predictive model results, normalization is performed on the continuous inputs to ensure that they are in a similar range (Lee, 2021).

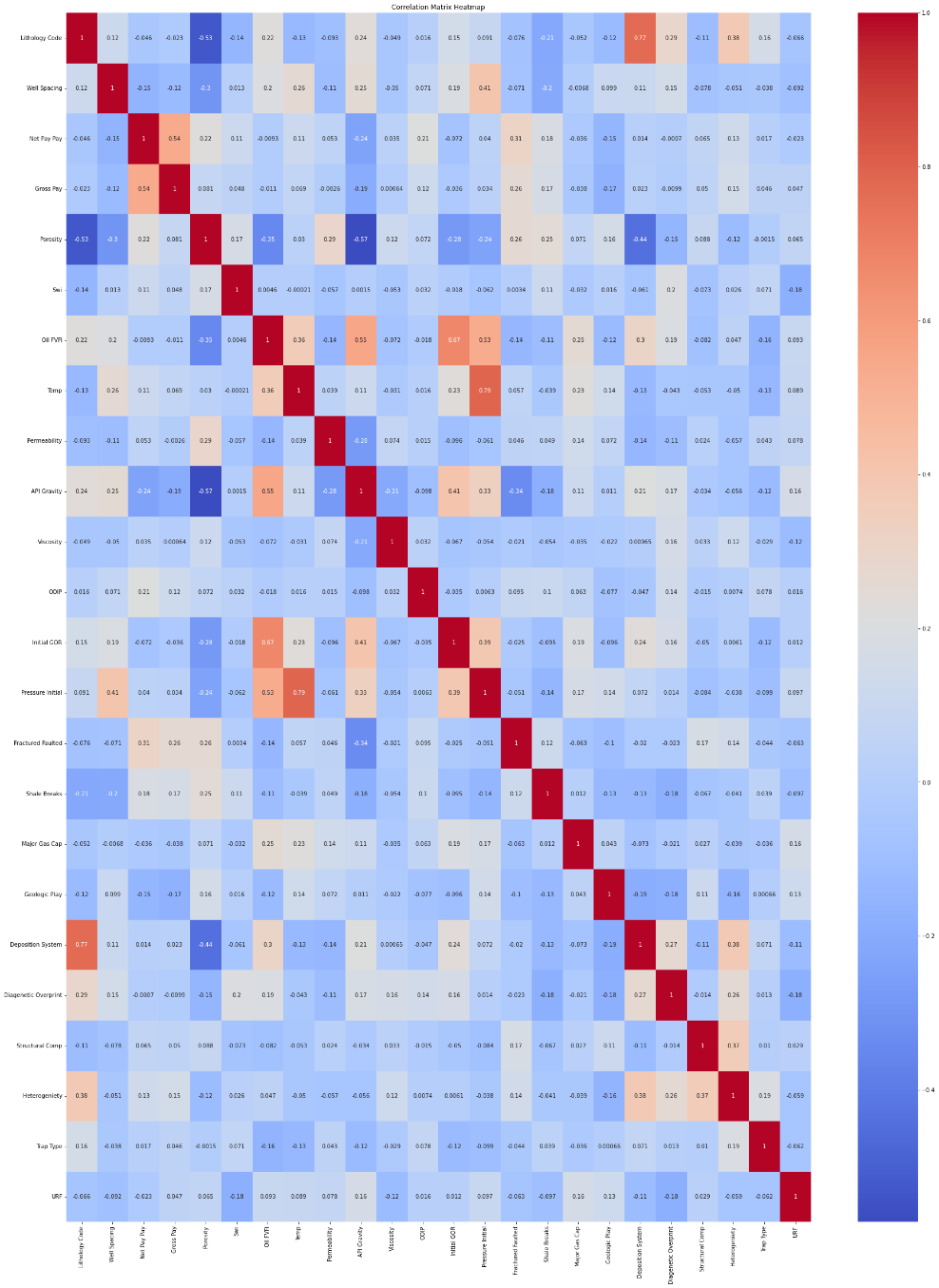
The TORIS dataset contains no missing values but it is observed that “Major Gas Cap” contains some unusual values.

**A graph with blue and black text

Description automatically generated**

**Figure 2. Plot of URF vs Major Gas Cap**

Based on figure 2, there are two odd values of 600 and 680 in the “Major Gas Cap” variable. Since it is supposed to be a flag variable where 0 = No Major Gas Cap and 1 = Has Major Gas Cap, the entire records where “Major Gas Cap” = 600 and 680 are removed instead of making assumptions to replace the values. Thus, the TORIS dataset now has 447 records.

****

**Figure 3. Correlation Matrix of TORIS Variables**

The Pearson Correlation is used where 0 indicates that the inputs are entirely not correlated and 1 implies that they are fully correlated. For this study, above 0.7 suggests high correlation. Based on figure 3, “Lithology Code” and “Deposition System” contain a correlation of 0.77 and “Pressure Initial” and “Temp” contain a correlation of 0.79. Despite both pairs being highly correlated, all inputs will be used during the initial phase of the construction of the ANN model and one variable from each pair may be removed subsequently to improve the results.

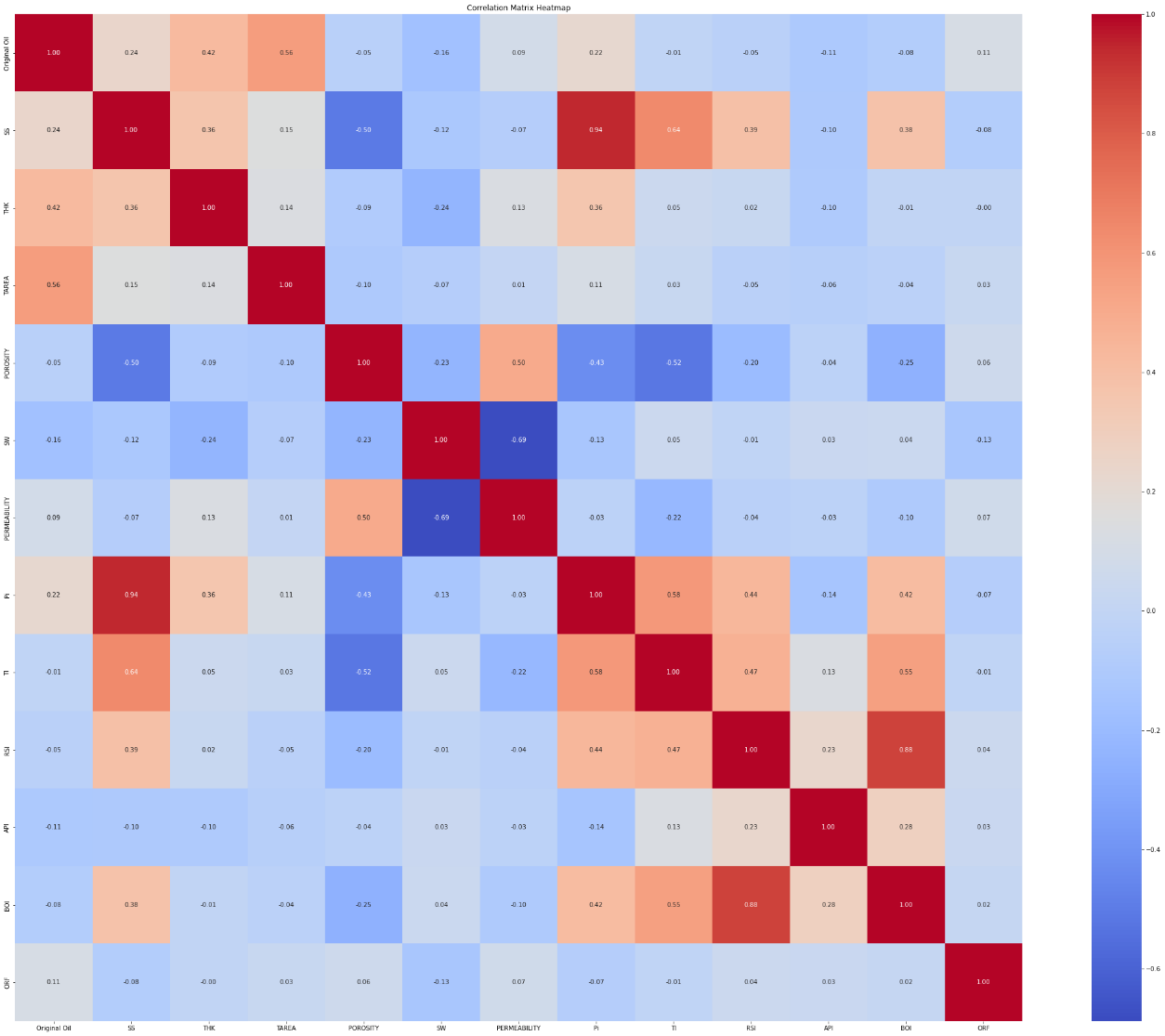
## GOM

A screenshot of a computer

Description automatically generated

**Figure 4. Statistics Table of GOM Variables**

For the GOM dataset, there are 2 categorical and 12 continuous inputs used to predict a continuous output “ORF” as shown in figure 4. There is no missing value in the GOM dataset. The dataset contains 4,512 records. Similarly, the continuous inputs have values of different ranges and normalization is used to transform the values. For the categorical variables, their values have also been changed due to label encoding. Previously, “CHRONOZONE” contained values such as PLL, PU, MUU, and 11 other values. The new values for “CHRONOZONE” are now from 1 to 14.



**Figure 5. Correlation Matrix of GOM Variables**

A correlation matrix is also performed on the GOM dataset. According to figure 5, two pairs are highly correlated with “Pi” and “SS” containing a correlation of 0.94 and “BOI” and “RSI” containing a correlation of 0.88. Similar to TORIS, all inputs will be used for the sake of constructing a simple ANN model during the first phase and hyperparameters will be tuned to improve the model’s performance.

# Proposed Modelling and Evaluation

ANN will be adopted for this project. The project will split into two phases as follows:

1. Construct ANN models using IBM SPSS Modeler with basic settings
2. Construct ANN models using Python and tune hyperparameters

During the initial phase, an ANN model with basic parameter settings will be constructed using IBM SPSS Modeler. However, constructing an ANN model with IBM SPSS Modeler is usually a black-box approach as it is unknown how the model derives its number of hidden layers and types of activation functions to achieve its optimized results. Moreover, tuning of hyperparameters is also more restricted with IBM SPSS Modeler compared to with Python (Chan et al., 2022). Due to the limitations, the second phase of this study will focus on using Python to construct a series of ANN models. This project will also analyze how the tuning of the various hyperparameters such as input selection, number of hidden layers, types of activation functions, number of epochs, and batch sizes will affect the ANN models’ results.

Subsequently, error rates like mean squared error (MSE) and MAE, and a statistic test of R², will used as evaluation techniques to compare the results of the ANN models constructed both in IBM SPSS Modeler and Python.

# Proposed Schedule

|  |  |
| --- | --- |
| Schedule | Task |
| Week 2-3  21 Aug 2023 (Zoom meeting with supervisor to prepare for business proposal submission) | * Work on proposal * Make amendments given by supervisor based on zoom meeting |
| Week 4  8 Sep 2023 (Business Proposal Submission 20%) | * Submit Business Proposal |
| Week 5-7  18 Sep 2023 (Zoom meeting with Supervisor to prepare for oral presentation) | * Construct basic ANN model on IBM SPSS Modeler and Python * Tune various hyperparameters on Python to obtain optimized ANN results * Make amendments given by supervisor based on zoom meeting |
| Week 8  2-6 Oct (Oral Presentation 20%) | * ANN Models to be in deployment stage * Complete Oral Presentation * Make amendments given by supervisor based on oral presentation |
| Week 9-12  30 Oct 2023 (Zoom meeting with supervisor to prepare for final report submission) | * Start working on final report * Make amendments given by supervisor based on zoom meeting |
| Week 13  6 Nov (Final Report Submission 60%) | * Submit Final Report |

# References

1. Afari, S. A., Kwame, S., Wilberforce, N. A., & Anthony, M. (2015). Assessment of petroleum reservoir recovery factor using complexity scoring and artificial neural network. *International Journal of Scientific Engineering and Research, 3*(7), 124-129.

2. Ahmed, A. A., Salaheldin, E., Chen, W., & Abdulazeez, A. (2019). Estimation of oil recovery factor for water drive sandy reservoirs through applications of artificial intelligence. *Energies*.

3. Bureau of Ocean Energy Management. (2019). *Atlas of Gulf of Mexico gas and oil sands data*. U.S Department of the Interior. https://www.data.boem.gov/Main/GandG.aspx

4. Chan, S. P., Tan, S. C., Wang, D., & Ren, J. (2022). ANL307 Predictive modelling (study

guide). *Singapore: Singapore University of Social Sciences*.

5. Gowtham, T., & Wu, X. (2017). Using data analytics on dimensionless numbers to predict the ultimate recovery factors for different drive mechanisms of Gulf of Mexico oil fields.

6. Kumar, M., Swaminathan, K., Rusli, A., & Thomas-Hy, A. (2022). Applying data analytics & machine learning methods for recovery factor prediction and uncertainty modelling.

7. Lee, K. H. C. (2021). ANL303 Fundamentals of data mining (study guide). *Singapore: Singapore University of Social Sciences*.

8. Li, H., Cao, N., Tian, H., & Cheng, S. (2020). Applications of artificial intelligence in oil and gas development. *Archives of Computational Methods in Engineering, 28*, 937-949

9. Sachin, C., & Karmakar, G. P. (2020). Artificial intelligence techniques and their application in oil and gas industry. *Artificial Intelligence Review*.

10. Shams, K., Sidqi, A. A., Hasan, Y. A., & Rahul, G. (2020). A novel empirical correlation for waterflooding performance prediction in stratified reservoirs using artificial intelligence. *Neural Computing and Applications.*

11. Surajudeen, S., Hassan, S., Afza, S., Raji, I. O., & Yarima, M. H. (2022). Prediction of electromagnetic properties using artificial neural networks for oil recovery factors. *Colloid Journal, 85*(1), 151-165

12. U.S Department of Energy. (1995). *Tertiary oil recovery information system.* U.S Department of Energy. https://edx.netl.doe.gov/dataset/2006-oil-and-gas-industry-software