

MACHINE LEARNING PREDICTION OF NMR POROSITY IN CARBONATE RESERVOIRS

BY

HUSSAIN ALI ALHAMOOD

A Thesis Presented to the

COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

PETROLEUM ENGINEERING

JUNE, 2024

**MACHINE LEARNING PREDICTION OF NMR POROSITY IN
CARBONATE RESERVOIRS**

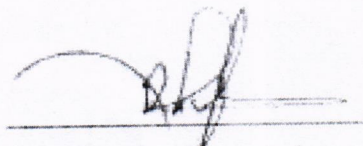
HUSSAIN ALI ALHAMOOD

PETROLEUM ENGINEERING

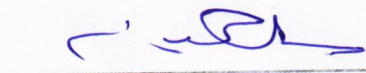
June 2024

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN- 31261, SAUDI ARABIA
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Hussain Ali Alhamood** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of: **MASTER OF SCIENCE IN PETROLEUM ENGINEERING**.



Dr. Dhafer Al-Shehri
Department Chairman



Dr. Suliman Saleh Al-Homidan
Dean of Graduate Studies

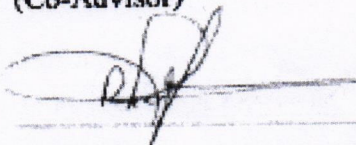
Date _____



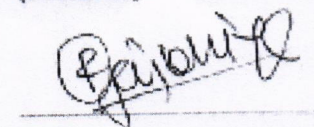
Dr. Mohammed Naser-Aldeen
Mahmoud
(Advisor)



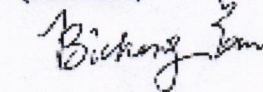
Dr. Abdulaziz Abdulrahim
(Co-Advisor)



Dr. Dhafer Al-Shehri
(Member)



Dr. Rahul N. Gajbhiye
(Member)



Dr. Bicheng Yan
(Member)



© Hussain Ali Alhamood

2024

Dedication

This work is dedicated to my parents, my wife, my daughter, my sister and my brothers. They have been always standing by my side during this journey, and I could not have completed this work without their support.

My father and mother, who have always believed in me and this achievement is small part of their effort. My wife, Noor, who always encourages me to seek my goals in life. My daughter Rahaf is the light the gives me hope in the difficult times. My sister Zahra, my brothers Mohammed and Muayad are the hands that lift me up when I need it most.

I would not forget the continuous encouragement of my friends and colleagues.

For all of this, thank you very much.

ACKNOWLEDGMENTS

I would like to send my sincere appreciation to Prof. Mohammed Naser-Aldeen Mahmoud for his support and guidance during the master degree. In addition, I would like to thank the committee members, Dr. Abdulaziz Abdulrahim, Dr. Dhafer Al-Shehri, Dr. Rahul N. Gajbhiye, and Dr. Bicheng Yan, for their great contribution to complete this work.

I extend my sincere thanks to the teaching faculty and the members of the College of Petroleum Engineering and Geosciences for their continuous support.

A special note of thanks to Dr. Ali Al-Turki, whose encouragement has been a motivation behind this accomplishment. |

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS	VI
LIST OF TABLES.....	IX
LIST OF FIGURES.....	XI
ABSTRACT	XIV
ملخص الرسالة	XVI
CHAPTER 1 INTRODUCTION.....	1
1.1 Problem Statement.....	1
1.2 Research Objective	1
1.3 Significance of the Study	2
1.4 Scope of the Research.....	2
1.5 Carbonate Rocks	3
1.5.1 Porosity.....	3
1.5.2 Total Porosity.....	5
1.5.3 The Pore System in Carbonate rocks	5
1.5.4 Porosity Measurements Methods	7
a. Porosity Measurement from Core Samples	7
b. Porosity Measurement from Well Logs.....	8
CHAPTER 2 LITERATURE REVIEW	15
2.1 Literature Review of Porosity Prediction	15
2.1.1 Timur Empirical Correlation, 1972	15
2.1.2 Magnetic Resonance Imaging Logging (MRIL) by Miller et al., 1990	16
2.1.3 Combinable Magnetic Resonance (CMR) by Chang et al., 1994	17
2.1.4 NMR from Pore Size Distribution by Prammer, 1994	17
2.1.5 Integration of NMR and Conventional Logs by Georgi et al., 1999.....	18
2.2 Machine Learning in Petroleum Engineering	19

2.2.1 Prediction of Porosity using Hybrid Computational Intelligence Models by Anifowose et al., 2010	19
2.2.2 Predict the Nuclear Magnetic Resonance Porosity of the Carbonate Reservoir by Mustafa et al., 2022	20
2.2.3 Machine Learning Based Prediction of Porosity from Varg Field by Andersen et al., 2022	20
2.2.4 Machine Learning for Estimating Petrophysical Properties of Carbonate Rocks Using NMR Core Measurements by Farmanov et al., 2023	21
CHAPTER 3 METHODOLOGY.....	22
3.1 Introduction to Artificial Intelligence (AI)	22
3.2 Machine Learning	22
3.2.1 Training in Machine Learning	23
3.2.2 Clustering Algorithms	30
3.2.3 Hyperparameter Tuning	36
CHAPTER 4 DATA ANALYSIS AND REVIEW	37
4.1 Data Description.....	37
4.2 Exploratory Data Analysis.....	40
4.2.1 Statistical Analysis	40
CHAPTER 5 RESULTS.....	75
5.1 Error Metrics	75
5.2 Silhouette Coefficient.....	77
5.3 Results.....	78
5.3.1 Primary Models (Model-A)	78
5.3.2 Optimized Models (Model-B)	87
5.3.2 Models with Clustered Data (Model-C)	93
5.3.4 Optimized Models with Clustered Data (Model-D)	101
CHAPTER 6 DISCUSSION AND CONCLUSION.....	114
6.1 Interpretation of the Findings	114
6.2 Limitations of the Study	119
6.2.1 Shale Impact on NMR Signals	119
6.2.2 NMR Analysis Discrepancy from LWD	119
6.2.3 NMR Analysis in Gas Hydrate-Bearing Sections.....	120
6.2.4 Limitation of Machine Learning Models	120

6.3 Conclusion	121
REFERENCES.....	123
VITAE	132

LIST OF TABLES

Table 1: Choquette and Pray classification for pore systems in carbonate rocks.	6
Table 2: Porosity measurement methods, their advantages, and disadvantages.	7
Table 3: Well logs type, their requirements, advantages, and disadvantages.	13
Table 4: Statics of the variables in data set number 1.	42
Table 5: Statics of the variables in data set number 1 after corrections.	42
Table 6: Statics of the variables in data set number 2.	43
Table 7: Statics of the additional variables in data set number 2.	44
Table 8: Statics of the lithology volumes in data set number 2.	44
Table 9: Statics of the variables in the compiled data set.	45
Table 10: Hypothesis testing results for the variables in data set number 1 with respect to total NMR porosity.	48
Table 11: Hypothesis testing results for the variables in data set number 2 with respect to total NMR porosity.	49
Table 12: Hypothesis testing results for the variables in the compiled data set with respect to total NMR porosity.	50
Table 13: Pearson's correlation coefficient calculations for the variables in data set number 1 with respect to total NMR porosity.	53
Table 14: Pearson's correlation coefficient calculations for the variables in data set number 2 with respect to total NMR porosity.	54
Table 15: Pearson's correlation coefficient calculations for the variables in the compiled data set with respect to total NMR porosity.	54
Table 16: The results of the machine learning models of data set number 1.	78
Table 17: The results of the machine learning models of data set number 2 (without the variables of zero readings).	81
Table 18: The results of the machine learning models of data set number 2 (without the lithology readings).	83
Table 19: The results of the machine learning models of the compiled data set.	85
Table 20: The results of the machine learning models of data set number 1 from the optimized model.	87
Table 21: The results of the machine learning models of data set number 2 (without the lithology variables).	90
Table 22:: The results of the machine learning models of the compiled data set.	92
Table 23: The results of the clustering algorithms using data set number 1.	94
Table 24: The results of the machine learning models of data set number 1 after adding the clustering features.	95
Table 25: The results of the clustering algorithms using data set number 2.	97
Table 26: The results of the machine learning models of data set number 2 after adding the clustering features.	98
Table 27: The results of the clustering algorithms using data set number 1.	100

Table 28: The results of the machine learning models of the complied data set after adding the clustering features.....	100
Table 29: The results of the machine learning models of data set number 1 (including the clustering features) in a new optimized model.....	102
Table 30: The results of the machine learning models of data set number 2 (including the clustering features) in a new optimized model.....	104
Table 31: The results of the machine learning models of the complied data set (including the clustering features) in a new optimized model.....	106
Table 32: The cross-validation results for Random Forest Regressor and Gradient Boosting Regressor algorithms.	108

LIST OF FIGURES

Figure 1: Porosity definition demonstration	4
Figure 2: Examples of Choquette and Pray classification for pore systems in carbonate rocks.	7
Figure 3: Total and effective porosity measurement methods [101].	14
Figure 4: Simple ANN model structure	24
Figure 5: Simple (MLP) ANN model structure	25
Figure 6: Components of Decision Tree model.	26
Figure 7: Simple Radom Forest Model.	27
Figure 8: The relationship of models' error and number of iterations using Gradient Boosting [55].	28
Figure 9: K-means data clustering at different iterations [104].	32
Figure 10: A simplified example of DBSCAN data points clustering [106].	33
Figure 11: Assigning exemplars and modifying the selection in Affinity Propagation clustering [40].	34
Figure 12: Lithology measurements for data set number 2	39
Figure 13: Percentage for the volumes of mineral in data set number 2.	45
Figure 14: Comparison of Pearson's correlation coefficient for the variables from the three data sets with respect to total NMR porosity.	55
Figure 15: Pearson's correlation coefficient for the variables the additional log measurements from data set number 2.	56
Figure 16: Pearson's correlation coefficient for the variables the lithology log measurements from data set number 2.	56
Figure 17: Pair plot for the correlations between the variables in data set number 1.	58
Figure 18: Heatmap for the correlations between the variables in data set number 1.	59
Figure 19: Histogram for data distribution of the logs' variables in data set number 1. ..	60
Figure 20: Pair plot for correlations between the logs' variables in data set number 2.	63
Figure 21: Pair plot for correlations between the additional variables in set number 2. ..	64
Figure 22: Pair plot for correlations between the lithology volumes in set number 2.	65
Figure 23: Heatmap for the correlations between the variables in data set number 2.	66
Figure 24: Histogram for the data distribution of the log variables in data set number 2.	67
Figure 25: Histogram for data distribution of additional logs' variables in set number 2.	68
Figure 26: Histogram for data distribution of lithology volumes in set number 2.	69
Figure 27: Histogram for data distribution of lithology volumes in set number 2.	70
Figure 28: Pair plot for correlations between logs' variables in the compiled set.	71
Figure 29: Heatmap for the correlations between the variables in the compiled data set.	72
Figure 30: Histogram for data distribution of logs' variables in the compiled set.	73
Figure 31: Error evaluation vs model complexity of training and testing groups [60].	77
Figure 32: Accuracy comparison of training and testing results for the primary machine learning models of data set number 1.	79

Figure 33: Accuracy comparison of training and testing results for primary machine learning models of data set number 2 (without variables of zero readings). ..	81
Figure 34: Accuracy comparison of training and testing results for the primary machine learning models of data set number 2 (without the lithology readings).	84
Figure 35: Accuracy comparison of training and testing results for the primary machine learning models of the compiled data set.....	86
Figure 36: Accuracy comparison of training and testing results for the optimized machine learning models of data set number 1.	88
Figure 37: Accuracy comparison of training and testing results for the optimized machine learning models of data set number 2 (without the lithology variables).....	90
Figure 38: Accuracy comparison of training and testing results for the optimized machine learning models of the compiled data set.....	92
Figure 39: Accuracy comparison of training and testing results for the machine learning models that include clustering features of data set 1.	96
Figure 40: Accuracy comparison of training and testing results for the machine learning models that include clustering features of data set 1.	98
Figure 41: Accuracy comparison of training and testing results for the machine learning models that include clustering features of the compiled data set.....	101
Figure 42: Accuracy comparison of training and testing results for the optimized machine learning models after including clustering features with set number 2.	105
Figure 43: Accuracy comparison of training and testing results for the optimized machine learning models after including clustering features with the compiled set... ..	106
Figure 44: The correlation between the predicted values and the actual values of the total NMR porosity using the Random Forest Regressor algorithm for the training and testing groups of data set 1.....	109
Figure 45: Distribution of actual and predicted values of total NMR porosity using Random Forest Regressor algorithm for testing group of set 1.	109
Figure 46: The correlation between the predicted values and the actual values of the total NMR porosity using the Gradient Boosting Regressor algorithm for the training and testing group of data set 1.....	110
Figure 47: Distribution of actual and predicted values of total NMR porosity using Gradient Boosting Regressor algorithm for testing group of data set 1.	110
Figure 48: The correlation between the predicted values and the actual values of the total NMR porosity using the Random Forest Regressor algorithm for the testing group of the compiled data set.	111
Figure 49: The distribution of the actual and the predicted values of the total NMR porosity using Random Forest Regressor algorithm for the testing group of the compiled data set.....	111

Figure 50: The correlation between the predicted values and the actual values of the total NMR porosity using the Gradient Boosting Regressor algorithm for the testing group of the compiled data set.	112
Figure 51: The distribution of the actual and the predicted values of the total NMR porosity using Gradient Boosting Regressor algorithm for the testing group of the compiled data set.	112
Figure 52: A simplified workflow for the used step to predict the total NMR porosity using machine learning.	118

ABSTRACT

Full Name : Hussain Ali Alhamood
Thesis Title : Machine Learning Prediction of NMR Porosity in Carbonate Reservoirs
Major Field : [Petroleum Engineering
Date of Degree : [June 2024]

Porosity is an important petrophysical measurement that provides general evaluation to the hydrocarbon reservoirs. The number of studies for porosity prediction is an indication for its importance. Major percentage of porosity in sedimentary rocks exist in carbonate rocks. The pore system in carbonate rocks is very complex due to the existence of different pore sizes and rock geometry. This complexity creates challenges in the prediction of porosity in carbonate reservoirs. The common data sources for porosity prediction are mainly core samples and well logs data. The comparison between the well logs data illustrates a clear advantage for NMR tool in porosity prediction over the other porosity tools as NMR porosity is independent of lithology of the rock formation. NMR porosity tool principle is the estimation of the existing hydrogen atoms from the fluids in the pore space.

This research is applying machine learning techniques to predict the total porosity estimated from Nuclear Magnetic Resonance (NMR) tool. 19735 data points from conventional well logs were used as the original set of data for this research. The data set is extracted from numerous wells that were drilled in an oil field located in the Middle East and the covered zone of the logs is identified to be from carbonate reservoir. Initial and focused data evaluation was conducted to assess the quality of the data and eliminate the outliers and invalid data points. A total of ten (10) machine learning algorithms were used to develop several machine learning models for total NMR porosity prediction. Additional stages of tuning and optimizing to the machine learning models were applied to enhance the results. The developed machine learning models' evaluation criteria are correlation coefficient (R), average absolute percentage error (AAPE), and root mean square error (RMSE).

The two machine learning algorithms random forest regressor and gradient boosting regressor delivered high accuracy result for the prediction of NMR porosity. The developed models for the algorithm have delivered prediction accuracy that reaches 97% with minor perdition error values. This solution can provide an alternative technique for the prediction of porosity in carbonate rocks and reduce the dependency on the traditional method of porosity prediction. Additional outcome of the study is a simplified workflow for utilizing machine learning algorithm to predict the total NMR porosity specifically and the petrophysical parameters in general.

This research supports the utilization of machine learning models to predict NMR porosity. Moreover, it is an indication for the validity of using machine learning to estimate the petrophysical parameters in carbonate reservoirs and the petroleum industry in general.

ملخص الرسالة

الاسم الكامل: حسين علي حسين الحمود

عنوان الرسالة: تقدير مسامية الرنين المغناطيسي النووي في مكامن الصخور الكربونية باستخدام أساليب التعليم الآلي

التخصص: هندسة البترول

تاريخ الدرجة العلمية: يونيو 2024

مسامية الصخور من أهم الخصائص البتروفيزيائية التي تستخدم لتقدير كميات الهيدروكربون في مكامن النفط و الغاز، يتجلى ذلك في عدد الدراسات المنشورة التي تعمل على تقدير المسامية في المكامن، جزء كبير من المسامية في الصخور المترسبة تتواجد في الصخور الكربونية، لكن نظام المسامات في الصخور الكربونية يتميز بشدة التعقيد و ذلك لتنوع أحجام المسامات و اختلاف أشكال الصخور أثناء عملية الترسيب، مما يجعل تقدير المسامية في مكامن الصخور الكربونية صعبا جدا، هنالك عدة مصادر شائعة لتقدير المسامية مثل عينات الحفر الجوفية أو تسجيل القياسات للآبار، عند مقارنة هذه المصادر يتضح وجود فارق كبير في جودة عملية تقدير المسامية عند استخدام أداة الرنين المغناطيسي النووي وذلك لعدم اعتمادها على معرفة نوع الصخور في المكن، حيث أن أداة الرنين المغناطيسي تعتمد على حساب عدد ذرات الهيدروجين في الموائع الموجودة في الفراغات المسامية.

هذه الدراسة تعمل على استخدام أساليب التعليم الآلي لتقدير المسامية الكلية لأداة الرنين المغناطيسي النووي، حيث تم استخدام 19735 عينة من البيانات من سجلات القياسات التقليدية للآبار كمصدر رئيسي للبيانات في هذه الدراسة، هذه السجلات تم استخراجها من عدة آبار تم حفرها في أحد حقول النفط في منطقة الشرق الأوسط، حيث تم تصنيف الجزء المغطى بالسجلات على أن هو مكن للصخور الكربونية.

الخطوات الأولى من الدراسة ركزت على تحليل البيانات و إزالة المعلومات الغير صحيحة، بعد ذلك تم استخدام عشرة (10) خوارزميات من خوارزميات التعليم الآلي لتطوير نظام يعمل على تقدير المسامية الكلية لأداة الرنين المغناطيسي النووي، بالإضافة إلى تطبيق عدد من الخطوات لتحسين عميلة التعليم للحصول على نتائج أفضل، تم قياس جودة التعليم واختبار النتائج باستخدام عدة معايير منها معامل الارتباط، ومتوسط النسبة المئوية للخطأ المطلق، وجذر متوسط الخطأ المربع.

النتائج بينت أن اثنتان من خوارزميات التعليم الآلي وصلت لنتائج عالية من الدقة في تقدير المسامية من أداة الرنين المغناطيسي النووي، جودة النتائج وصلت إلى نسبة 97% من الدقة مع وجود نسبة خطأ بسيطة في تقدير المسامية، حيث بالإمكان استخدام هذه الوسيلة لحساب و تقدير المسامية الكلية في الصخور الكربونية كحل بديل للوسائل التقليدية

مما يقلل الاعتماد و الحاجة إلى هذه الوسائل، بالإضافة إلى ذلك، تم إنشاء طريقة عمل مبسطة لاستخدام خوارزميات التعليم الآلي لحساب المسامية بشكل خاص و الخصائص البتروفيزيائية بشكل عام.

هذا البحث يبين قدرة التعليم الآلي على حساب المسامية من أداة الرنين المغناطيسي النووي مما يشجع على استخدام هذه الوسيلة لحساب الخصائص البتروفيزيائية في صناعات النفط الغاز.

CHAPTER 1

INTRODUCTION

This chapter introduces the problem, the causes, and the motivation of the research. In addition, defines key terms and explains the methods used to conduct the research.

1.1 Problem Statement

Carbonate reservoirs are one of the major reservoirs in the petroleum industry, and understanding their properties and behavior is crucial for successful reservoir exploration and production. The prediction of porosity in carbonate reservoirs is challenging due to the variety of pore sizes and rock geometries. Nuclear Magnetic Resonance (NMR) logs can provide a solution for porosity prediction with accurate measurements as NMR porosity is lithology-independent. However, the high cost and time requirement of NMR logging often led to the selection of other, less accurate prediction methods.

1.2 Research Objective

This study aims to develop a machine learning model to predict NMR porosity in carbonate reservoirs from conventional well logs (gamma ray, neutron porosity, photoelectric log, and density logs) with high accuracy. This will provide an alternative, cost-effective, and time-efficient method for porosity prediction, eliminating the need for NMR logging.

1.3 Significance of the Study

Nuclear Magnetic Resonance (NMR) logs are widely recognized as the most accurate method for measuring porosity in carbonate reservoirs, providing invaluable reservoir data. However, their high cost and time-consuming nature limit their widespread application. Consequently, there is a growing need for alternative methods that offer comparable accuracy but are more cost-effective and time-efficient. Machine learning models trained to predict porosity from conventional well logs emerge as a promising solution, potentially replacing NMR logs and providing accurate measurements of total porosity in carbonate reservoirs.

1.4 Scope of the Research

The research included the following stages:

- Gather the necessary well data, which includes Gamma Ray, Caliper, Neutron Porosity, Density, Resistivity, and Sonic well logs. Additionally, lithology measurements might be included to verify the associated conditions of the reservoir.
- Prepare and re-arrange the collected data into a unified, tabular format suitable for machine learning model consumption.
- Conduct a comprehensive exploratory data analysis.
- Apply feature engineering analysis and classify the available features.
- Develop machine learning models to test the different methods.
- Evaluate the models and select best performing model based on pre-defined evaluation metrics.

1.5 Carbonate Rocks

Carbonate rocks are type of sedimentary rocks that are mainly composed of carbonate minerals. The types of the carbonate rocks are classified based on the contained minerals. The first type is Limestone, which is mainly composed of the carbonate minerals calcite or aragonite (CaCO_3). The second type of carbonate rocks is Dolomite rock, which is known as Dolostone, and the main component for this rock is the carbonate mineral dolomite [$\text{CaMg}(\text{CO}_3)_2$] [66]. Also, a mixture of fine-grained calcite with terrigenous mud creates Marl, which might contain aragonite and dolomite. Carbonates rocks are mainly formed as a result of biological and biochemical processes that include fossils accumulations and encrusting organisms. Also, some inorganic processes such as calcite precipitation from seawater can form calcite coating or cementation. This usually takes place in shallow warm marine environments, where the suitable conditions of temperature, pH, and ion concentrations are available [97].

1.5.1 Porosity

Porosity is the petrophysical parameter that measures the ratio of the void space volume to the total volume of the formation [Eq.1]. Different fluids including gas, oil, and water can occupy the void space of the formation. Formations with large void space volumes have the ability to store high volumes of fluids. This is an indication of the porosity measurements with the formation as the fluids' volumes are higher in the formations with high porosities. This means that porosity is a key parameter to quantify the number of stored hydrocarbons with the formations. Since hydrocarbon production is one of the main objectives for formation analysis, then porosity can be used as an indicator for formation

evaluation [42]. Besides that, the formation lithological facies can be identified and classified by using porosity data as porosity ranges varies based on the type of rocks. [54].

$$Porosity = \frac{Pore\ Volume}{Bulk\ Volume} \quad [Eq. 1]$$

$$\phi = \frac{V_P}{V_B} \quad [Eq. 2]$$

Where:

ϕ : Porosity

V_P : Pore Volume

V_B : Bulk Volume

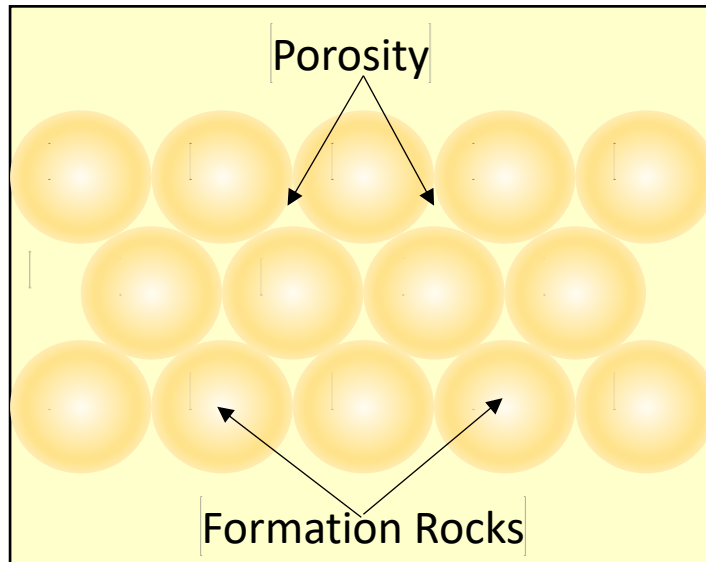


Figure 1: Porosity definition demonstration

1.5.2 Total Porosity

Porosity is classified based on the time of void space formulation within the rock into primary and secondary porosities. The primary porosity can be identified as the associated porosity with the rock deposition or formation. The secondary porosity is defined as the formed porosity after the process of the rock deposition such as vugular spaces and fractures. Porosity also classified based on the contents of the pore space of the rock into total and effective porosity. The total porosity is measured by calculating the ratio of the total pore space in the rock to the bulk volume of the rock [Eq.3]. The Effective porosity is calculated by subtracting the fraction of the pore space occupied by shale from the total porosity. So, only pore spaces that store hydrocarbons are included in the calculations of effective porosity.

$$\phi_t = \phi_e + V_{sh} \times \phi_{sh} \quad [\text{Eq. 3}]$$

Where:

ϕ_t : Total porosity, fraction

ϕ_e : Effective porosity, fraction

V_{sh} : Shale volume, fraction

ϕ_{sh} : Shale porosity, fraction

1.5.3 The Pore System in Carbonate rocks

The pore system in carbonate rocks is different than the pore system in other types of rocks, where it is more complex in carbonate rocks. The reason of this complexity is the existence of several pore sizes within the same rock as some carbonate rocks might exhibit variety pore sizes. Besides that, the observed geometry of the pores in carbonate rocks is variable. Moreover, the diagenesis processes such as precipitation, fracturing, and chemical

dissolution increases the complexity of carbonate rocks structures. The presence of these conditions leads to creating very heterogeneous pore system in carbonate rocks [42].

The pore systems in carbonate rocks are classified into three (3) basic groups based on Choquette and Pray classification, which are fabric selective, non-fabric selective, fabric selective or not [Table-1]. The fabric selective porosity includes interparticle, intraparticle, intercrystal, moldic, fenestral, shelter, and growth framework. The types of the non-fabric selective porosity are fractures, channels, vug, and cavern. The examples of the third group of porosity are breccia, boring, burrow, and shrinkage [Figure-2][37].

Generally, the porosity types of the fabric selective group are characterized as primary pore systems. On the other hand, the non-fabric selective pore systems are usually classified as secondary porosities [32].

Basic Porosity Type in Carbonate Rocks		
Fabric Selective Porosity	Non-Fabric Selective Porosity	Fabric Selective or not
Interparticle, Intraparticle, Intercrystal, Modlic, and Fenestral.	Fractures, Channels, Vug, and Carvern.	Breccia, Boring, Burrow, and Shrinkage.

Table 1: Choquette and Pray classification for pore systems in carbonate rocks.

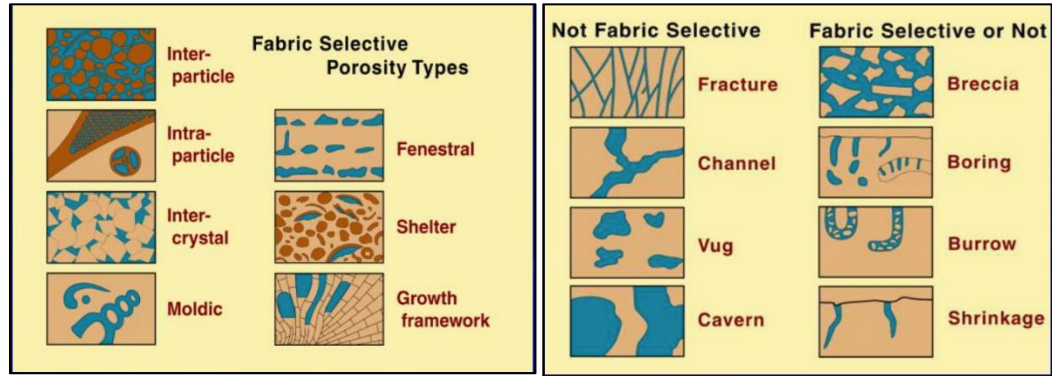


Figure 2: Examples of Choquette and Pray classification for pore systems in carbonate rocks.

“A New Genetic Classification of Carbonate Porosity and Its Application to Reservoir Characterization”
[64]

1.5.4 Porosity Measurements Methods

There are two common sources that are used to obtain the porosity data. These two sources are core samples and well log data. A summary for the advantages and disadvantages of the data sources is highlighted in table-2.

Method	Advantages	Disadvantages
Core Samples	Most Representative	- High cost - Difficult to maintain - Only cover short sections
Well Logs Data	Provides more data points	- Selecting the suitable correlation - Requires calibration with core data

Table 2: Porosity measurement methods, their advantages, and disadvantages.

a. Porosity Measurement from Core Samples

There are several methods to measure porosity from core samples. Some examples of these methods are Imbibition method, Mercury Injection method, and Gas Expansion method. These methods are usually taking place in the laboratory. To measure the porosity from core samples, the pore volume and the bulk volume of the core sample must be obtained.

This requires the determination of some associated values such as sample weight, test fluid density, or test fluid volume. The porosity results from core samples are considered to be accurate. These results are typically reflecting the measurements of total porosity only [52]. The utilization of core sample as the source of porosity data is limited due to the high cost of extracting the core samples. In addition, it is difficult to maintain the samples during the process of core extraction and transportation.

b. Porosity Measurement from Well Logs

Well logs are the most common source for porosity data as the well logging process covers large sections of the formation by a single log. Besides that, porosity data from well logs can be attained at several stages of the well life. There are several logs that can serve the objective of porosity data anticipation. These logs include density, sonic, neutron, and nuclear magnetic resonance (NMR).

i. Density Logs

The main objective of the density logs is to obtain the formation bulk density. In addition to the density measurement, the density logs are one of the well logging methods that can be utilized to measure formation porosity. The main principle of density log is the measurement of the reduced gamma ray that was sent from the tool source to formation. The gamma rays with an initial known value travel from the tool radioactive source to the tool detectors through the formation. During this path, the gamma rays are being reduced by the interaction with the electrons in the outer shell. The returned radiation to the tool is received either by the short-range detector or the long-range detector. The amount of the reduced gamma ray is reflecting the formation bulk density, where the high density causes

higher reduction of the gamma ray and low density shows low reduction of the gamma ray. This means the reading of the gamma ray is inversely proportional to the reading of the gamma ray at the receivers. From [Eq.4], the relationship between the formation bulk density, the matrix density, and the formation fluid density can be obtained. The total porosity can be acquired from the density measurements as showing in [Eq.5] [42/53].

$$\rho_b = (1 - \phi)\rho_{ma} + \phi\rho_f \quad [\text{Eq. 4}]$$

$$\phi_t = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_f} \quad [\text{Eq. 5}]$$

Where:

ρ_b : Formation bulk density

ϕ_t : Porosity

ρ_{ma} : Matrix density

ρ_f : Formation fluid density

ii. Neutron Porosity Logs

One of the methods that are used to estimate the porosity is the neutron porosity log. This tool is used to predict the liquid-filled porosity, where the tool responds mainly to hydrogen atoms in the formation. The concept of using the neutron porosity logs is sending high-energy neutrons to the formation from a radioactive source. The released neutrons spread into the formation and lose their energy by reacting with the hydrogen atoms. This leads to create high-energy gamma ray, which represent the number of the hydrogen atoms in the formation. As water and liquid hydrocarbons have the highest concentrations of hydrogen atoms, this means that the zones with high porosity will show a low record of gamma ray. Also, it will display a high record of gamma ray for the low porosity rocks. The cause of this behavior is the required distance through the formation for the neutrons to lose their energy by reacting with hydrogen. Moreover, the matrix rocks of the formation must be

defined before running the neutron porosity log as it is affected by the lithology of the formation and it should be recalibrated if it was measured for different type of rocks. In addition, the presence of gas in the formation leads to underestimate the porosity measurements, where gas has low hydrogen concentration. On the other hand, shale provides higher measurements of porosity due to the high volume of bond water within shale rocks [53].

iii. Sonic Porosity Logs

Sonic logs, which are known as acoustic log, measures the required time for an elastic wave to transfer from the transmitter to the receiver. The use of this tool aims to predict the velocity of the elastic wave through the formation, which will lead to estimate the porosity of the formation rocks. As the measurements of the wave travel time increases, the porosity of the rock's increases [53].

The porosity of the rocks can be predicted from the chart of the transit time verses porosity, where the resulted transit time of the sonic log tool will be used on the chart to determine porosity. The use of the chart requires selecting the type of the tested rock. Also, the transit time measurement is affected by the existence of gas as gas tend to slow the transfer of wave, which will lead to overestimate of porosity. In addition, the sonic waves only measure the primary porosity, where the sonic waves travel in the formation by selecting the fastest path through the rock matrix. This means it will avoid the fluid contents, fractures, and vuggs [9]. Moreover, the produced wave of the sonic log tool has higher frequency than the seismic wave. This could lead to different results of the applications of

the two methods [53].

iv. Nuclear Magnetic Resonance (NMR) Porosity

The Nuclear Magnetic Resonance (NMR) method is mainly used to predict porosity and pore size distribution in the formation. The concept of this method is applying a magnetic field on the hydrogen nuclei, where hydrogen exists abundantly in the fluids of the pore spaces including water and hydrocarbons. In addition, hydrogen atoms have an odd number of protons (^1H) and they can produce the measurable signal by the NMR tool. The applied magnetic field is stronger than the magnetic field of the earth by hundred times in order to create measurable signals. The detected amount of NMR signal is directly related the number of the magnetic moments that are responsible for generating the signal.

The process of the NMR well logging to predict porosity involves several stages, which are known as the CPMG (Carr-Purcell- Meiboom-Gill) sequence. These stages are divided based on the different rotation movements of the hydrogen protons. The two movements created by the protons are spin and precession. Generally, hydrogen protons are randomly aligned on different spin axes. At this stage, no measurable signal will be observed, as there is no external magnetic influence. By applying an external magnetic field ($\vec{B}_0$) from the NMR tool on the hydrogen nucleus, the protons will be aligned and spin with respect to the applied magnetic field. The NMR tool contains an antenna that sends Radio-Frequency (RF) pulses into formation. These pulses are usually electro-magnetic pulses and they are measured by Larmor frequency (ν_0). The generated field is called ($\vec{B}_1$) and it aims to create distribution of the aligned protons. The caused motion of the pulses is perpendicular (in

the z-axis) to the magnetic field ($\vec{B_0}$) and it is known as precession. By removing implanted RF pulses, the hydrogen protons will try to retain their equilibrium states. In order to describe the effect of the applied forces on the hydrogen atoms, (2) two relaxation mechanisms have been obtained. These two relaxation mechanisms are called T1 and T2, respectively. The T1 relaxation mechanism is used to estimate the required time to have all hydrogen protons magnetized (net magnetization) and aligned with the magnetic field ($\vec{B_0}$). The T2 relaxation mechanism measures the cumulative time for a series of echo decays during the phase-out of the applied magnetic field into the x-y plane. T1 and T2 relaxation times are depended on the fluid properties, formation rock matrix, and well structure. By estimating the relaxation mechanisms, the porosity of examined rocks can be interpreted [18].

v. Well Logs Comparison

Each type of these well logs work based on different principles and it has different requirements. [1]. Table-3 display the list of these well logs data, their requirements, the advantages, and disadvantages.

Well Log Type	Required Data	Advantages	Disadvantages
Density Logs	- Matrix density - Fluid density	Low impact in formation gas presence	- Shallow investigation depth - Impacted by wellbore washouts
Sonic Logs	- Matrix transit time - Fluid transit time	Well-compensated for environmental effects	Depth of investigation highly dependent on formation compactness
Neutron Porosity Logs	Lithology Calibration	- Used in cased hole - Detect formation gas presence	- Requires lithology calibration - Highly impacted by borehole conditions (size variation)
Nuclear Magnetic Resonance (NMR) Logs	Hydrogen index	Lithology Independent	- High cost - Requires long time - Might be impacted with speed run of the tool

Table 3: Well logs type, their requirements, advantages, and disadvantages.

In addition, figure-3 demonstrates the different usage of the measurement methods to estimate the total and effective porosity. Besides that, it highlights the limitations of these methods, where the measurements of total and effective porosity might be overestimated or underestimated due to certain conditions of pore sizes, type of fluids, or the formation rock types. These limitations can be resolved by using Nuclear Magnetic Resonance logs as they are lithology independent and they use hydrogen index to estimate the porosity. So, by using NMR tool, the capillary water, bound water, and structural water measurements can be estimated [101].

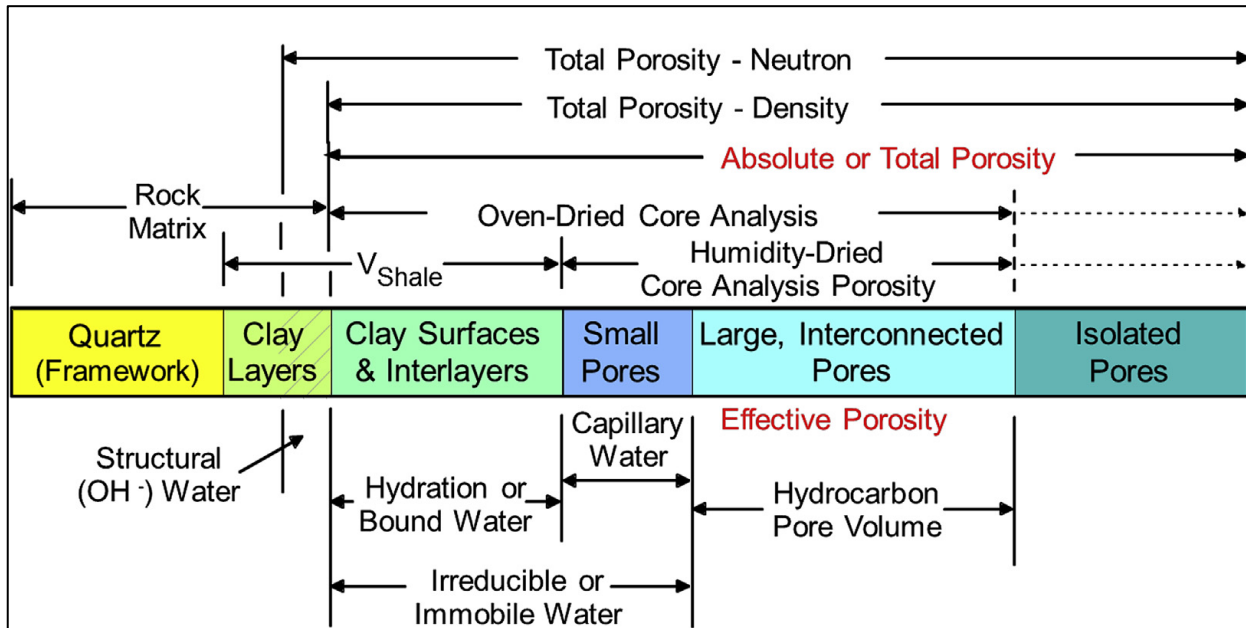


Figure 3: Total and effective porosity measurement methods [101].

This comparison between the well logs data demonstrates a clear advantage of NMR tool in porosity prediction for carbonate rocks over the other well logs tools. The reason for this preferability is the indecency of NMR porosity measurements from the lithology of the formation rocks. Although NMR logs are the favorable method in prediction of porosity in carbonate reservoir, it cannot be used in the porosity measurements for all wells. This due to associated high costs and the long time for NMR logging procedures.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature Review of Porosity Prediction

During the last decade, there were many studies for porosity measurement. The studies were discussing different factors that affect the porosity such as the formation type, environmental conditions, source of data, and analysis tools. Since the research methods have enhanced with the innovation of new technologies, the study of porosity prediction has been improved significantly. Below are some of these studies and their focus research area.

2.1.1 Timur Empirical Correlation, 1972

Timur research was established based on the estimation of the pore space volumes that are occupied by the moveable fluids, which are the fluids that can be extracted from the reservoir by applying the required production mechanisms. The research was investigating the relationship between the producible porosity (ϕ_p) and the free fluid index (FFI), which is measured by using spin-echo nuclear magnetic log (NML) apparatus in the lab. The initial study was applied in laboratory for sandstone samples and it provided good matching between the producible porosity (ϕ_p) and the free fluid index (FFI) in percent of bulk volume. After that, the developed model was tested on carbonate rocks samples with different pore space volumes in order to measure the total porosities of the rocks. The carbonate rock samples consisted from divergent ratios of limestone and dolomite, but they did not exhibit any major clay volumes. Similarly, the FFI for the carbonate samples has

been estimated by using NMR data and it was utilized to predict the total porosity. The experiment led to the empirical equation [Eq.6], which was used to compare the measured porosity from the rock samples with the predicted porosity from FFI. The outcomes of these analysis delivered excellent matching, but it cannot apply to all pore sizes. The limitation of this method is the underestimation of porosity in very small pores, which provides relaxation time reading of less than 15 msec. Besides that, these results have been obtained from laboratory experiments at specified conditions and it only confirms the results for a small selected sample of data [95/96].

$$\phi = 1.4 FFI + 0.3 \quad \text{[Eq. 6]}$$

2.1.2 Magnetic Resonance Imaging Logging (MRIL) by Miller et al., 1990

In 1990, Miller et al, have presented the magnetic resonance imaging logging (MRIL) tool. The objective of tool testing was delivering on site analysis for NMR logs. Also, the tool was able to provide continues reading NMR relaxation times directly from the borehole. The tool reduced to the requirements of formation lithology analysis as it provides lithology independent porosity measurements. The results of testing MRIL was compared to the porosity measurements from core samples and conventional logs including density and neutron porosity logs for sandstone and limestone. The outcome numbers of porosity were providing good match with other measurements methods with minor differences in sandstones sections. This research has confirmed the capability of using NMR signals in porosity prediction in different formation types. A constrain the need to be considered is not receiving the signals from clay, which does not produce measurable MRIL signals [68].

2.1.3 Combinable Magnetic Resonance (CMR) by Chang et al., 1994

The combinable magnetic resonance (CMR) tool was introduced in 1994 by Chang et al. The tool was tested in order to overcome the issue of measuring reservoir parameters in carbonate formation from resistivity logs. The challenge of carbonate reservoir was mainly caused by the variety of pore sizes, where it ranges from microporosity to vuggy porosity. Moreover, the lithology identification limits the capability of conventional logs in very heterogenous reservoirs. The experiment of the CMR tool was conducted on the porosity range of 0 to 25 p. u. in very complex carbonate formation, where it consists of a texture range from mudstone dolomite to grainstone dolomite with the existence of anhydrite and silt layers across the formation. The porosity measurements from the CMR tool were compared with NMR porosity from core samples analysis in the lab. The results are provided corresponding plots between the two methods with 1.4 p.u. higher measurements for NMR porosity. Besides that, the CMR was not able to identify the microporosity in the intervals of slit that provides less 1 msec reading for the relaxation time T2. This means it only provide the measurement of effective porosity in these intervals and cannot detect the zones of residual oil in the formation [30].

2.1.4 NMR from Pore Size Distribution by Prammer, 1994

Prammer was aiming to expedite the process of NMR analysis to predict from NMR logs directly at the well site. He presented the idea of estimating the reservoir porosity and permeability from the distribution of the pore sizes in 1994. The principle of this method is the interpretation of T2 relaxation spectra from NMR logs, where the distribution of spectrum values over the pore spaces in the bulk volume was investigated. In this study,

the reading from the relaxation spectrum T2 values were suggested to be averaged over the total pore sizes of the bulk volume. The test was conducted under the assumption that the pore spaces are filled with single fluid with uniform diffusion effects in the pore spaces, where the test is only considering water wettability. The research provided appropriate results where the controlled conditions are applied, but it exhibits major results discrepancy in the sections that contain more than one fluid. Moreover, this method cannot be implemented with areas of high content or the formations with significant pore size distribution. This means that the method cannot be utilized in carbonate reservoirs, where the pore systems is considered very complex [76].

2.1.5 Integration of NMR and Conventional Logs by Georgi et al., 1999

The evaluation of the reservoir petrophysical parameters by the integration of NMR data and conventional logs was conducted by this study in 1999. The main objective of the study was to reduce the dependency on core analysis in NMR logs interpretation and improve formation evaluation of the shaly sand zones. In addition, the research highlighted the impact of the reservoir conditions on the decay profile of NMR T2, where NMR logs are lithology independent but it can be changed by the pore fluid viscosity and diffusivity. An initial analysis for the T2 decays was used to distinguish the effective and total porosity based on the used range of NMR T2 profile. In addition, the study stated the classification for the effective porosity by using T2 cutoff, where it is divided to bulk volume irreducible water (BVI) and bulk volume free fluid (BVM). The integration with neutron density logs managed to predict the total porosity and resolve the issue of missing NMR data interpreted for clay bound water. A comparison with core sample analysis indicates the NMR

integration with conventional logs can assist in the process of reservoir characterization by correcting the NMR data due fluid properties effects [50].

2.2 Machine Learning in Petroleum Engineering

There are many challenges in the petroleum industry that requires the utilization of machine learning. In recent years, the implementation of machine learning in petroleum applications has increased. The number of AI published studies, which have been completed by the petroleum engineers, highlights the importance of machine learning in oil and gas fields. The studies aim to provide quick solutions by applying simple methods with high accurate results. Some examples of these studies are reservoir simulations, production prediction, and reservoir characterizations [3]. Several studies were targeting the prediction of porosity using the machine learning approach. Below are some examples for the researches that discussed the prediction of porosity.

2.2.1 Prediction of Porosity using Hybrid Computational Intelligence Models by Anifowose et al., 2010

This research was aiming to predict the properties of porosity and permeability in oil and gas reservoirs with high accuracy using. The approach that was used in this research is the combination of two or more machine learning models with different machine learning techniques. This approach is known as Hybrid Computational Intelligence (HCI). The objective of this approach was to enhance the prediction capabilities of the models and provide highly accurate results. The machine leaning models that were tested in this paper was the combination of Fuzzy Logic (FL) type-2, Support Vector Machine (SVM), and

Functional Networks (FN). The input data in the research for the porosity prediction are well log data from drilling site in North America. The best results for the testing of this study have exceeded 0.94 for the correlation coefficient (R^2) and 2.80 for the root mean square error (RMSE). The study demonstrated the capabilities of an individual case of high computational assets. Besides that, the study did not include the conditions of the reservoir and the type of rocks utilized for the porosity prediction [17].

2.2.2 Predict the Nuclear Magnetic Resonance Porosity of the Carbonate Reservoir by Mustafa et al., 2022

The paper is discussing the results of predicting NMR porosity in carbonate rocks using machine learning method. It illustrated the challenges in measuring porosity in carbonate reservoirs by traditional methods and demonstrates the advantages of NMR porosity. It suggested the utilization of conventional well logs data such as Gamma Ray, Sonic log, and neutron porosity in machine learning model to predict NMR porosity. Different machine learning algorithms were tested in this paper such as ANN and ANFIS. The model has used more 3500 data points and produced the result of 0.94 for the correlation coefficient (R^2), 0.39 for the root mean squared error (RMSE), and 5.09% for the average absolute percentage error (AAPE). More enhancements to the study results will lead to more sustainable method to predict NMR porosity, more reliable approach to predict porosity in carbonate reservoirs [19].

2.2.3 Machine Learning Based Prediction of Porosity from Varg Field by Andersen et al., 2022

This study was evaluating the porosity prediction in Varg field, Central North Sea, using

machine learning. The applied machine learning technique for this research was Least Square Support Vector Machine (LSSVM). The study has used the well log data including Gamma Ray, self-potential, sonic, bulk density, neutron porosity, deep and medium resistivity, and caliper logs as the inputs of the model. The study highlighted that the evaluated data are obtained from sandstone reservoir. The split of the data set in this research was (90%) for the training data and (10%) for the testing data. The high percentage for the training data set might lead memorizing the data values within the model. This might lead to overfitting status, where the results of the model are overestimated. The provided results from the research are 0.79 for the correlation coefficient [15].

2.2.4 Machine Learning for Estimating Petrophysical Properties of Carbonate Rocks Using NMR Core Measurements by Farmanov et al., 2023

The objective of this research was generating correlations to predict the petrophysical properties including porosity, permeability, and irreducible water saturation in carbonate reservoirs. The concept of the study is the comparison of NMR measurements and core lab results. Several machine learning methods were tested in this study including decision trees, ANN, and random forest. There were two sub-cases within the study, where T2 spectrum was used as input in one case and feature extraction was used in the second case to obtain the model's inputs. The results for the porosity prediction by deep neural network were showing the correlation coefficient (R^2) of 0.87 for the first case and 0.74 for the second case. The study depends on the availability of core measurements and NMR log data. There might be lack of accessibility to these sources of data in many cases of reservoir petrophysical analysis, which is considered as an additional challenge. Moreover, further analyses are required to deliver more improved results [44].

CHAPTER 3

METHODOLOGY

3.1 Introduction to Artificial Intelligence (AI)

Artificial intelligence can be defined as the ability of digital machines and computer systems to perform and handle tasks that requires human judgment and involvement. Generally, the objective of AI studies involves problem solving, reasoning, planning, classification, and natural language processing. For this purpose, AI has been tested and implemented in many industries such as medicine, marketing, transportation, education, and other industries that can be optimized by using AI [35].

3.2 Machine Learning

Machine Learning is the science of the Artificial Intelligence (AI), where the machines (computers) utilize the data in order to find a solution for a suggested problem. On the contrary of normal programming, the computers in machine learning models are enabled to use algorithms and statistical models to learn and complete certain tasks without direct programming. There are several machine learning algorithms that can be applied to handle different types of problems. For example, the machine learning models can be applied in order for classification, prediction and regression problems [51].

3.2.1 Training in Machine Learning

a. Artificial Neural Network (ANN)

Artificial neural network (ANN) is one of the machine learning tools that provide the capabilities of solving complex and scientific problems. The general principle of ANN is dividing multipart data structures into minor sets of data, where each data set is handled separately to provide a specific solution and results will be combined to produce a cumulative solution for a problem. Typically, ANN contains three (3) layers, which are the input, output, and hidden layer. Each layer is consisted of smaller units that are called neurons, which are connected to the neurons of other layers by specifying the function of connection. Besides that, a weight for each input parameter is assigned to evaluate its impact in the network. The weight of each parameter will be adjusted by passing each node (neuron). The objective of the weight adjustment is to match the model output values with the target values by minimizing their differences after passing each node. The weight adjustment is conducted by applying mathematical activation functions. Also, there are functions that aims to identify the best group of model parameters that helps to match the target values, which are known as solver functions. Moreover, other model parameters that can be used to enhance the model capabilities to predict the target values such as number of hidden layers, model learning rate, and the maximum number of iterations in the model. Figure-3 demonstrates the structure of a simple ANN model [72].

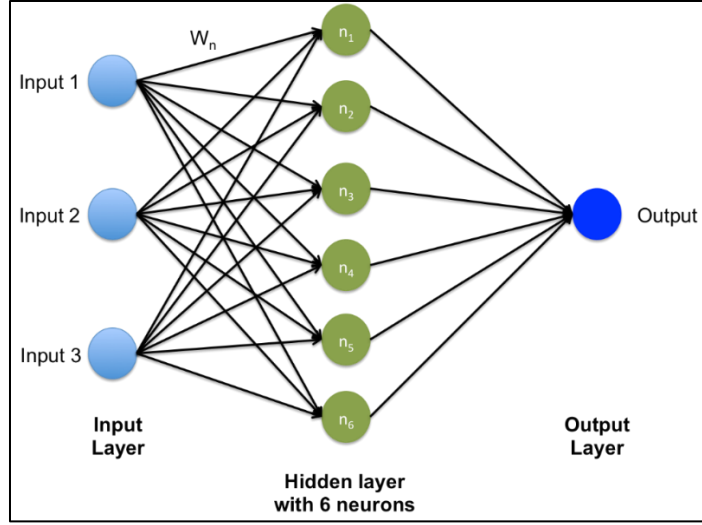


Figure 4: Simple ANN model structure

i. Multi-Layer Perceptron (MLP) Regressor

Multi-Layer Perceptron (MLP) Regressor is one of the ANN common types that is used to predict continuous output variables (regression problems). Similar to ANN, MLP contains three (3) input, output, and hidden layers. The architecture of MLP usually transfers the signals in one direction of the network from the input layers towards the output layers passing through the hidden layers. This architecture is known as feedforward neural network, which means that assigned weight for each parameter will be adjusted at each node (neurons) without looping any output from the nodes. Thus, the values will not go through the previous layer of the network again. There are many activation functions and solver functions that can be used to train within the MLP models. Some examples of the activation functions are Rectified Linear Unit (ReLU), Hyperbolic Tangent Function (tanh), and Sigmoid Function (Logistic). Examples of the solver functions that can be applied within the MLP model are Stochastic Gradient Descent (SGD), Adam, and LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno). [75]. In figure-4, a simple structure of MLP model is illustrated.

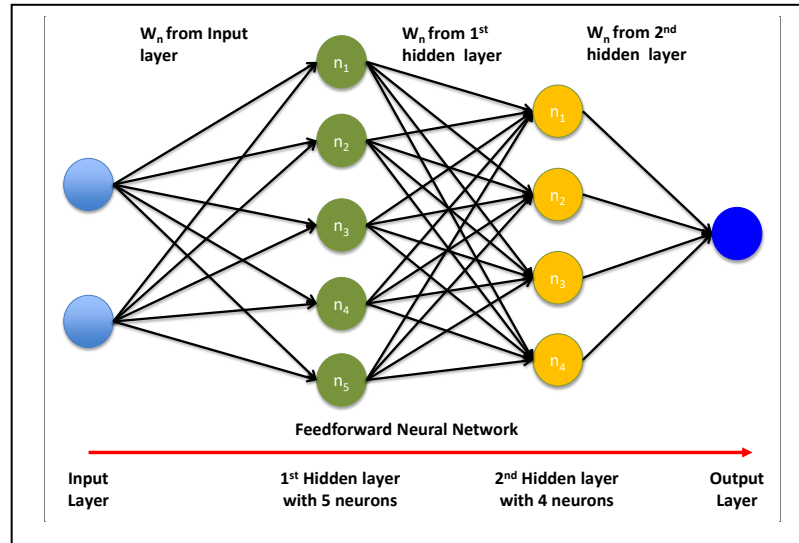


Figure 5: Simple (MLP) ANN model structure

b. Decision Trees

Decision tree is a type of the supervised of machine learning methods that can be utilized to find solutions for classification (discrete output values) and regression (continuous output values) problems. This method links the features (model's inputs) to target variable (model's outputs) by splitting the features into set of nodes based on answers ("True" or "False") for set of questions. The questions are created by the model using information from the model's inputs. The answers of the questions provide decisions to the models. These decisions from the first set of questions will be used as inputs to a second different set of nodes to create new decisions. There are model parameters that administrate the decisions from each set of questions such as splitter, specifies the splitting mechanism, and criterion that select the function to evaluate the quality of the splitting [103].

The splitting steps of will be repeated by the model to reduce the differences between the model's inputs and outputs to predict the nearby results. Repeating these steps will create a model structure similar to the tree structure. For this reason, the decision tress models are

using terms like Root Nodes, Decision Nodes, Branches, and Leaf Nodes to describe the different parts of the model [77]. The components of a decision tree model are displayed in figure-5.

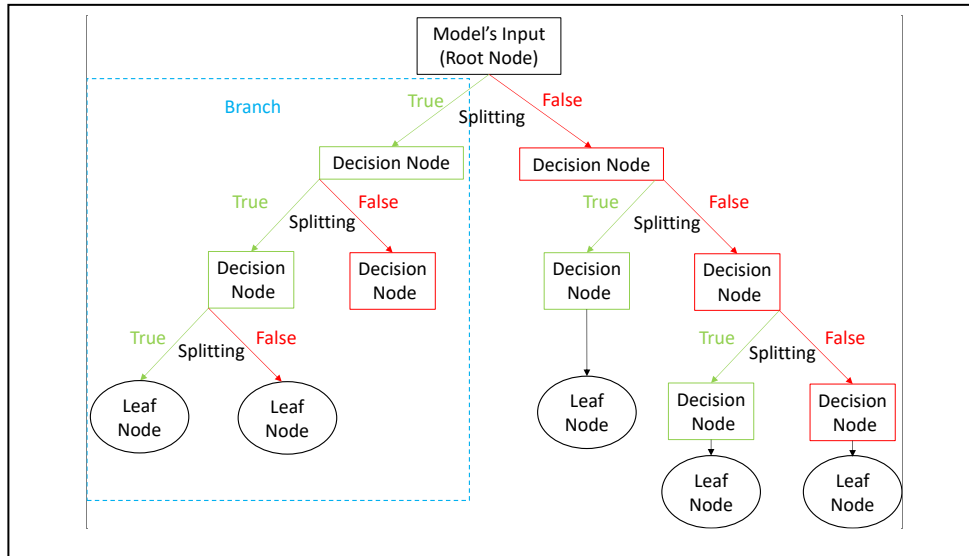


Figure 6: Components of Decision Tree model.

c. Random Forest

Random Forest is a machine learning method that was developed based on the concept of decision trees. Random forest creates several decision trees within the random forest model in order complete the training. Normalization will be applied to the results of these decision trees and single result will be provided from each decision tree. This normalization step will contribute to the elimination of overestimating the model results (overfitting), which is a limitation in the decision tree method. The random forest method can be applied to predict a class for classification problems or a value for regression problems. The single result (output) in the classification problems will be the mode of classes. On the other hand, the mean average will be the results of the forest random model in regression problems [24/ 84]. Figure-6 demonstrate the extension of decision trees to construct a random forest model.

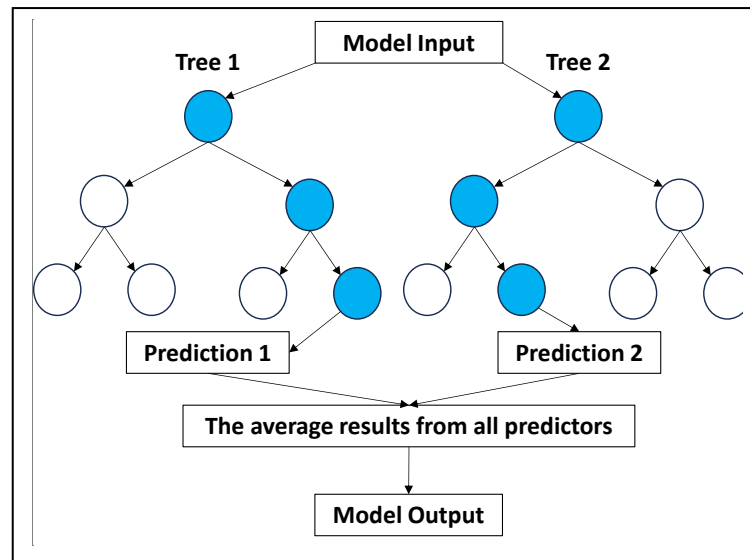


Figure 7: Simple Random Forest Model.

d. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning method. SVM method aims to create a hyperplane dimensional space with the N number of dimensions, where N is the number of the model's inputs (features). This will lead to classify the data points into separate groups based on selected kernel functions such as sigmoid, radial basis function, and polynomial kernel functions. The SVM model will try to allocate the hyperplane that provides the biggest differences between the groups of data points. The ability to separate data points into different classes indicates to the effectiveness SVM in classification (categorization) problems as it can create a hyperplane space with large number of dimensions. Although SVM is very effective in classification problems, but it can be used to predict values from regression problems [7/ 17].

e. Gradient Boosting

One of the machine learning techniques is Gradient Boosting. The Gradient Boosting

method helps to enhance the machine learning models' accuracy and prediction capabilities. The concept of the Gradient Boosting is improving the performance of the machine learning model by adding new models to enhance the prediction of regression or classification problems. The new models, which are commonly decision tree models, work to reduce the resulted bias errors from the initial models. By repeating this process, the Gradient Boosting algorithm can achieve better prediction results and low model errors from the first models, where the predictions from all the models are combined and each model is weighted according to its results. In addition, some training factors might be applied to adjust the models' weights during the training process. One of the properties of using Gradient Boosting algorithm in machine learning models is minimizing the overfitting of models results, where several regularizations are implemented within the model [49]. Figure-7 demonstrates the relationship between reducing the models' error with the number of iterations by using the gradient boosting technique.

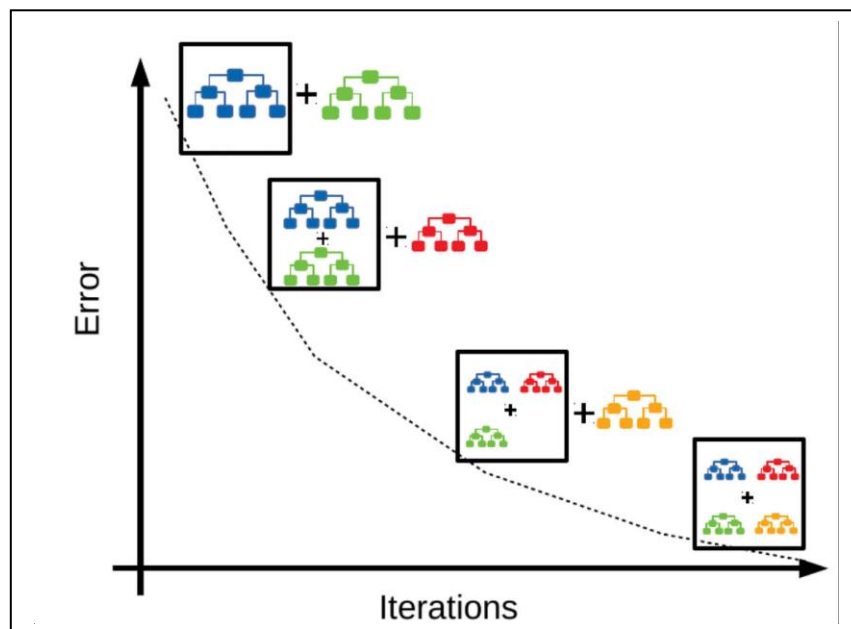


Figure 8: The relationship of models' error and number of iterations using Gradient Boosting [55].

i. eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting (XGBoost) is an optimized machine learning algorithm of the gradient boosting technique. The objective of using this algorithm is to enhance the machine learning model speed and capabilities to find the solution of regression and classification problems. The initial machine learning models used in XGBoost algorithm are typically decision trees. One of the benefits of using XGBoost algorithm from the common Gradient Boosting technique is reducing the results overfitting, where several techniques are implemented with XGBoost algorithm. In addition, XGBoost algorithm can provide more flexibility to change algorithm parameters as well as validating the model results at each iteration (cross-validation). Moreover, the XGBoost machine learning algorithm can expand the investigation of positive results into deeper models of decision trees than Gradient Boosting algorithm [48].

ii. eXtreme Gradient Boosting Random Forest Regressor (XGBRFRegressor)

Another algorithm that was developed based on the Gradient Boosting technique is eXtreme Gradient Boosting Random Forest Regressor (XGBRFRegressor). The XGBRFRegressor applies the same concept of Gradient Boosting technique and includes the specifications of random forest algorithm. So, the initial models in this machine learning approach will be decision trees, but will be trained using the random forest algorithm. This combination can be used to handle data sets with many features to enhance the results and reduce the overfitting results [102].

iii. Light Gradient Boosting Machine Regressor (LGBMRegressor)

Light Gradient Boosting Machine Regressor (LGBMRegressor) is another algorithm that implements the Gradient Boosting technique. This algorithm was developed to overcome the challenges of handling big data sets for machine learning studies. This algorithm is the favorite choice in the situation of limited computing resources, where it can add speed and efficiency to the learning process. The LGBMRegressor can deliver high accuracy results as it has smaller splitting range (leaf-wise) and deeper investigation depth. This could lead to overfitting results in some cases, where the splitting range might be too small for splitting [47].

3.2.2 Clustering Algorithms

The clustering analysis is one of the used methods to improve the results of the machine learning models. The process of clustering the data aims to classify and distribute the data into groups based on the similarities of the different data points. By combining the similar data points into one group, the process of identifying the group will be easier as the model will handle the data based on the group level instead of the data point level. This means that the model will analyze the data that contain similar structures and patterns as one component, which will ease the process of identifying the data [92]. In addition, the clustering analysis will improve the performance and accuracy of the machine learning process as it will help in expending the learning process and eliminating some choices within the data identification. Moreover, the number of groups for the clustered data, which known as the number of clusters, can be used as additional feature (input variables) for the machine learning model [71]. There are several clustering algorithms that can be used to

classify the data. Each of these algorithms apply different concept in separating the data points into groups.

a. K-means Algorithm

The K-means algorithm is one the clustering algorithms that are used for data clustering as part of the machine learning analysis. The applied concept of clustering the data in K-means algorithm is classifying the data into selected number of clusters, which are defined as K clusters. The number of clusters must be selected before clustering the data points. The number of clusters represents the number of centroids. The centroid is the mean distance between the data points in each group. By implementing the K-means algorithm for one time on the data points, each data point is assigned into the nearest centroid. The group of data points that have been linked to a specific centroid are identified as one cluster. After that, the process will be repeated by allocating new centroids, which represents the new mean distances. This will lead to re-assigning the data points to the new nearest centroids, which will update the number of data points in each cluster. In order to approach the most suitable classification for data points into nearest centroid, the process of re-assigning the data points is repeated until no change in the classification. In figure-8, an example of clustering the data at the initial cluster selection and the changes of clusters after several iterations using K-means algorithm is demonstrated. There is an issue with using K-means algorithm to cluster the data points, where the initial selection for the number of clusters is arbitrary. By choosing a different number of clusters, the clustering results will be different. Besides that, K-means algorithm is not sufficient for outliers' elimination, which impact the clustering process. [92].

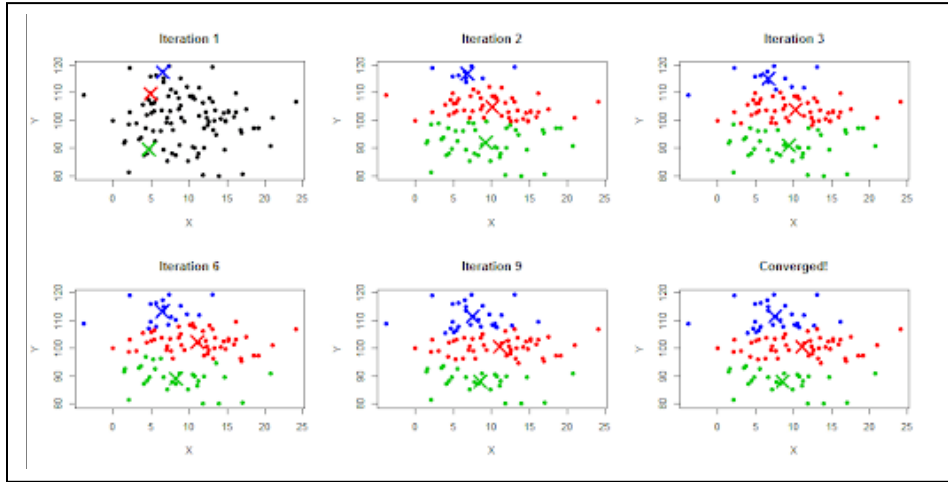


Figure 9: K-means data clustering at different iterations [104].

b. DBSCAN Algorithm

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is another algorithm that can be used for data clustering prior to the machine learning process. This algorithm works to separate the different data points into clusters based on density. This means that the data points are separated into groups based on the number of data points, where the groups of high number of data points are highlighted as dense areas. The centers of the clusters in DBSCAN are data points that are characterized by specified criteria. These points are highlighted as core points. The core points must have a minimum number of data points (including itself) within the specified distance to be labeled as core points. The minimum number of points is defined by the *min_samples* parameter and the distance are defined by the *eps* parameter. The points within a neighborhood of a core point are known as boarder points. In the case of having a data point that does not apply the criteria of core point nor considered as boarder point, it be labeled as noise point. This means it is a cluster outlier and it will be eliminated, which is considered an advantage for DBSCAN clustering over other algorithms such as K-means. Figure-9 provides a simple

example for clustering the point based DBSCAN algorithm. By repeating the process to with unlabeled data points, new core points, boarder points, and noise points will be highlighted. The DBSCAN clustering will be completed once no more points are added to core points or boarder points. Similar to K-means algorithm, the number of clusters must be selected to distribute the data among the clusters. In addition, this technique might not be accurate once handling data points with many features (input variables) as the investigated range of each feature is different from the others [41].

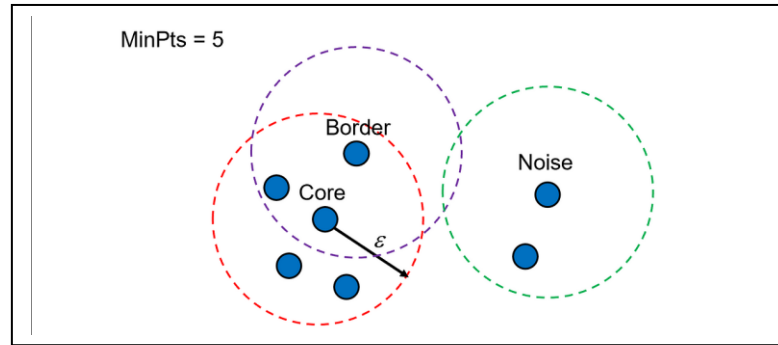


Figure 10: A simplified example of DBSCAN data points clustering [106].

c. Affinity Propagation Algorithm

Affinity Propagation Algorithm is unique method of data clustering. Unlike K-means and DBSCAN algorithms, Affinity Propagation Algorithm does not require a specification for the number of clusters. The Affinity algorithm will initially select different data points to represent the clusters. These data points will be called exemplars. After that, the concept of exchanging messages between data points to evaluate the quality of the selected exemplars to represent the cluster. With iterating the process, the data points with the best quality of being exemplars will be identified and the number of clusters will be decided. Figure-10 display the different stages of selecting the exemplars and unselecting with more iterations. Affinity Propagation Algorithm can be very sufficient is data clustering in some

cases, but it might not deliver the best clustering results in other cases as the algorithm is very sensitive to the parameter selection. Besides that, it will require high computing resources to evaluate the exemplars selection in large sets of data points [45].

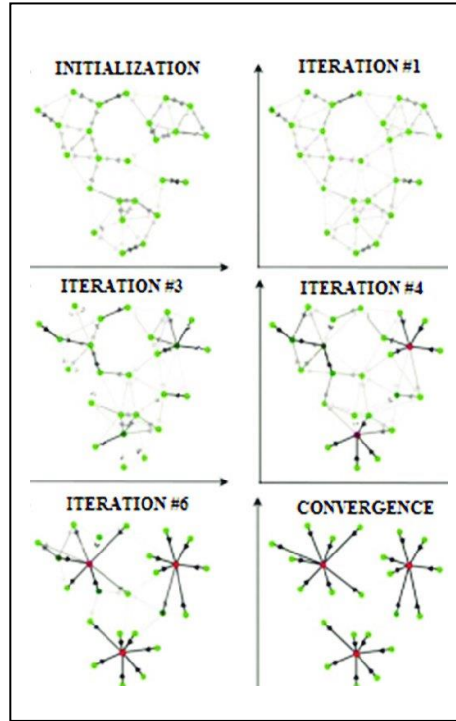


Figure 11: Assigning exemplars and modifying the selection in Affinity Propagation clustering [40].

d. BIRCH Algorithm

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) can overcome the issue of clustering large sets of data. In order to handle data with many features, BIRCH applies the conception of Clustering Feature (CF), Subcluster Features (SF), and CF tree, where CF is a small part of the data that provides a summary of the whole data set and SF are smaller parts of the clusters. This means that BIRCH depends on some statistical calculations to represent the cluster. These calculations include the number of data points in the cluster (N), the linear sum of the points (LS), and the sum of the squares of the points

(SS). The first step of the BIRCH clustering is to initiate the CF tree with hierarchical structure, where each node of the tree is representing a cluster. After that, the CFs and SFs will be injected to the CF tree and the tree structure will be evaluated. In order to maintain a stable tree structure, nodes might be split into new node or merged into one node. This process will be under a continuous evaluation for the clusters to check for the possibility of clusters merging or the need of creating clusters. The clustering process will be completed once all data points are assigned into a cluster. This method of data clustering requires accurate selection of algorithm parameters as they might highly impact the clustering results [56/94].

e. Mean Shift Algorithm

Another algorithm that is used for data clustering is the Mean Shift algorithm. This algorithm tries to allocate the areas of higher density from the data points. The beginning of the clustering process, the algorithm handles each data point as a cluster. This means that the number of clusters is not required as input parameter for the algorithm, which is considered as advantage over some algorithms. The density of each cluster (data point) will be evaluated using *Kernel Density Estimation* function. The objective of this step is to determine how dense the accumulation of data points with regards to the selected clusters centroids. After allocating the centroid of each cluster, the cluster centroid will be shifted in the direction of areas with higher densities. In order to determine the magnitude and the direction of the movement, the mean shift vector for the data points is calculated. The calculations will measure the mean of differences between the data points and their neighbors of points based on the evaluation of Kernel Density function. These steps will

be repeated and the data points will continue their shift until the convergence stage, where the data points show no movement in their positions. Once the convergence stage is achieved, the number of the clusters will be determined based on the points distributions. Determining the most suitable number of clusters can be achieved efficiently by using the mean shift algorithm. On the other hand, the iteration of shifting the data points might be suitable for small data sets, but time consuming and requires highly computing resources for large data sets [34].

3.2.3 Hyperparameter Tuning

The machine learning models usually consist of multiple parameters and hyper-parameters. The parameters are the internal variables within the machine learning models, which are being modified during the models' training in order to reduce the error values between the actual the predicted values. Some examples of the model parameters are the weights and biases [22]. On the other hand, hyperparameters are constant variables of the training algorithm that are set before the beginning of the training stage in order to control the training process of the machine learning model. This means that the hyperparameters are not impacted by the data itself, but selecting the most suitable hyperparameters can improve the performance of the machine learning models. Each training algorithm has different set of hyperparameters. Changes in hyperparameters might increase or decrease the performance of the machine learning models. In order to achieve the best results of the models, the best combination of hyperparameters must be selected. The process of choosing the best hyperparameters is known as hyperparameters tuning or model optimization [79].

CHAPTER 4

DATA ANALYSIS AND REVIEW

4.1 Data Description

Two sets of conventional well logs were used the original set of data for this study. The first data set contains 14045 data points and the second data set contains 5802 data points. Both data sets are extracted from numerous wells that were drilled in an oil field located in the Middle East. Below is a brief description for each data the two sets.

- Data Set number 1:

The first data set consist of 14045 data points (112 data points were removed due to incorrect readings of Gamma Ray). The covered zone of the logs is identified to be from carbonate reservoir. The set of data includes depth, gamma ray, caliper, neutron porosity, photoelectric, density, and nuclear magnetic resonance logs.

- Data Set Number 2:

The second data set includes 5802 data points. This set of data has additional information than the first data set. Beside the measurements of depth, gamma ray, caliper, neutron porosity, photoelectric, density, and nuclear magnetic resonance logs, the second data set has the readings for the followings:

- i. Resistivity logs:
 - a. Deep Laterolog (LLD).
 - b. Shallow Laterolog (LLS).
 - c. Mean Spherical focused Log (MSFL)

- ii. Gamma Ray:
 - a. Potassium (POTA).
 - b. Uranium (URAN).
 - c. Thorium (THOR).

- iii. Sonic log:
 - a. Compressional wave traveltime (DTC).
 - b. Shear wave traveltime (DTS).

- iv. Lithology:
 - a. Anhydrite volume (AHYD)
 - b. Calcite Volume (CALC)
 - c. Dolomite Volume (Dolomite)
 - d. Chlorite Volume (FECHLORITEVOLUM)
 - e. Gypsum (GPSM)
 - f. Illite Volume (Illitevolume)
 - g. K-Feldspar Volume (KFEL)
 - h. Montmorillonite Volume
 - i. Pyrite Volume (PYRT)
 - j. Quartz Volume (QRTZ)

The lithology data points in second set were utilized to identify the reservoir rock type. As displayed in figure-12, the presence of calcite through the whole tested zone is an indication of carbonate formations. This is supported by several thin layers of dolomite that are observed in the tested zone. The availability of quartz indicates to the fine-grain size of the rocks. In addition, there is a major presence of montmorillonite and illite, which suggests the potential for shale or mudstone. The combination of these minerals with the inspected proportions leads to the classification of sand-rich carbonate rocks for the reservoir.

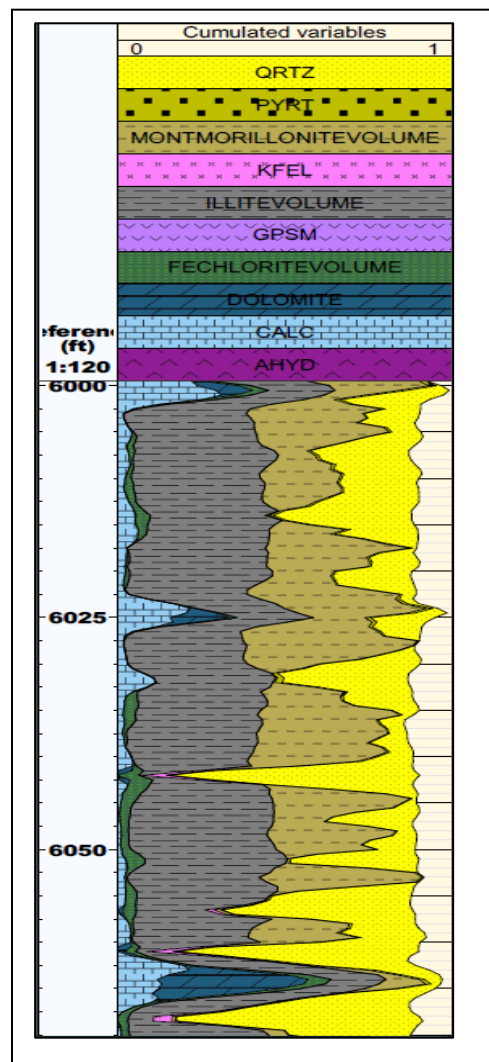


Figure 12: Lithology measurements for data set number 2

- Compiled Data Set:

A third data set was created by compiling the measurements for the similar features (log measurements) from the two data sets. The compiled data set includes 19735 data points of depth, gamma ray, caliper, neutron porosity, photoelectric, density, and nuclear magnetic resonance logs. The objective of compiling the data points from the two data sets is provide more training and testing data points for the machine learning models. This will help to improve the machine learning model performance, reduce the data variances, and create stable and reliable models.

4.2 Exploratory Data Analysis

The data exploration is an important and crucial part of the data analysis for the machine learning development process. It highlights the main characteristics of datasets and reduces the incorrect decisions and assumptions. In addition, it assists the process of initial models' development and allocating the relationship between the models' variables. All of this will enhance the machine learning model performance and provide better prediction results from the models [85].

4.2.1 Statistical Analysis

Statistical analysis is an essential part of the data exploration. It involves several steps such as data gathering, analyzing data, and allocating similar patterns between the data. Statistical analysis can help to transfer the raw data into structured set of data, which can

provide conclusions from the data. The tables below (table 4 to table-9) show the statistical analysis for the three data sets [67/82].

- Data Set number 1:

Table-4 displays different measurements for the variables in data set number 1 that includes the range, maximum, minimum, mean, mode, median, standard deviation, kurtosis, and skewness. The table shows the range of (0.20 to 14.113%) for the NMR total porosity. Besides that, it is observed that the mode for gamma ray reading is (-999.25), which has been showing for 112 data points (less than 0.8% of the total data points). The reading for these data points is invalid and they require more data review. Table-5 includes the updated numbers after removing the incorrect reading of gamma ray. Comparing the values in the two tables indicates that the changes for the data set measurements after removing the incorrect data is minor. The overall ranges for the variables in table-## are considered normal and realistic for carbonate reservoirs.

Index	Depth	Gamma Ray	Caliper	Neutron Porosity	Photoelectric	Density	Total NMR Porosity
Count	14045	14045	14045	14045	14045	14045	14045
Minimum	5417.00	-999.25	7.89	0.03	2.52	2.43	0.20
Maximum	8928.00	222.36	9.10	0.26	5.41	2.85	14.11
Mean	7172.50	92.77	8.86	0.17	3.24	2.59	6.16
Mode	N/A	-999.25	8.83	0.18	3.14	2.63	8.43
Median	7172.50	105.11	8.85	0.18	3.13	2.58	6.15
Standard Deviation	1013.65	100.24	0.08	0.03	0.41	0.06	2.24
25%	6294.75	92.15	8.82	0.16	2.98	2.55	4.59
50%	7172.50	105.11	8.85	0.18	3.13	2.58	6.15
75%	8050.25	113.84	8.90	0.19	3.35	2.63	7.69
Kurtosis	-1.20	109.38	23.95	2.21	4.54	1.11	-0.23
Skewness	0.00	-10.31	-2.72	-1.29	1.94	0.70	0.12

Table 4: Statics of the variables in data set number 1.

Index	Depth	Gamma Ray	Caliper	Neutron Porosity	Photoelectric	Density	Total NMR Porosity
Count	13933	13933	13933	13933	13933	13933	13933
Minimum	5417.00	14.65	7.89	0.03	2.52	2.43	0.20
Maximum	8900.00	222.36	9.10	0.26	5.41	2.85	14.11
Mean	7158.50	101.54	8.86	0.17	3.24	2.59	6.18
Mode	N/A	101.26	8.83	0.18	3.14	2.63	8.43
Median	7158.50	105.26	8.85	0.18	3.13	2.58	6.17
Standard Deviation	1005.56	21.57	0.08	0.03	0.42	0.06	2.23
25%	6287.75	92.66	8.82	0.16	2.98	2.55	4.61
50%	7158.50	105.26	8.85	0.18	3.13	2.58	6.17
75%	8029.25	113.90	8.90	0.19	3.35	2.63	7.70
Kurtosis	-1.20	2.61	26.22	2.20	4.49	1.12	-0.22
Skewness	0.00	-0.74	-2.84	-1.29	1.93	0.70	0.11

Table 5: Statics of the variables in data set number 1 after corrections.

- Data Set Number 2:

Similar to table-5, table-6 provides the same statistical measurements for data set number 2. By comparing the two tables, it is observed that the two data sets have a lot of similarities in the different statistical measurement parameters. This indicates to the reasonable ranges for the variables in data set number 2 as in data set number 1. Besides that, the statistical measurements have been obtained for the additional variables in data set number 2, which can be seen in table-7. Moreover, the statistical measurements for the lithology volumes from data set number 2 is calculated in table-8. In addition, figure-11 displays the percentage for the volumes of minerals in data set number 2.

Index	Depth	Gamma Ray	Caliper	Neutron Porosity	Photoelectric	Density	Total NMR Porosity
Count	5802	5802	5802	5802	5802	5802	5802
Minimum	5999.50	14.65	7.89	0.03	2.52	2.43	0.21
Maximum	8900.00	222.36	9.10	0.26	5.41	2.85	13.99
Mean	7449.75	102.83	8.85	0.17	3.27	2.59	5.92
Mode	N/A	101.56	8.85	0.18	3.14	2.58	5.99
Median	7449.75	106.97	8.84	0.18	3.16	2.59	5.88
Standard Deviation	837.52	22.46	0.08	0.03	0.44	0.06	2.22
25%	6724.63	94.50	8.81	0.16	3.00	2.55	4.38
50%	7449.75	106.97	8.84	0.18	3.16	2.59	5.88
75%	8174.88	115.31	8.89	0.19	3.39	2.63	7.39
Kurtosis	-1.20	2.68	28.13	2.36	3.83	0.82	-0.15
Skewness	0.00	-0.79	-3.01	-1.35	1.84	0.54	0.21

Table 6: Statics of the variables in data set number 2.

Index	Deep Laterolog (LLD)	Shallow Laterolog (LLS)	Mean Spherical focused Log (MSFL)	Potassium (POTA)	Uranium (URAN)	Thorium (THOR)	Compressional wave traveltime (DTC)	Shear wave traveltime (DTS)
Count	5802	5802	5802	5802	5802	5802	5802	5802
Minimum	4.28	3.97	0.24	0.32	1.19	0.65	52.10	185.57
Maximum	5184.96	4573.02	20000.00	4.19	15.33	15.82	92.73	200.19
Mean	39.27	38.47	39.69	2.31	9.59	4.37	80.18	192.75
Mode	8.73	12.14	20000.00	2.17	9.13	7.25	77.39	194.07
Median	16.43	14.49	8.65	2.38	9.83	4.10	80.45	192.81
Standard Deviation	176.69	179.48	599.60	0.58	2.45	1.53	5.75	2.02
25%	10.56	9.54	5.62	1.97	8.13	3.45	77.57	191.31
50%	16.43	14.49	8.65	2.38	9.83	4.10	80.45	192.81
75%	29.07	25.34	14.28	2.75	11.35	5.00	84.12	194.21
Kurtosis	400.77	342.26	1059.87	0.31	0.21	6.11	2.44	-0.06
Skewness	18.40	17.02	32.14	-0.66	-0.57	1.74	-1.10	-0.12

Table 7: Statics of the additional variables in data set number 2.

Index	AHYD	CALC	DOLOMITE	FECHLORITE	GPSM	ILLITE	KFEL	MONTMOR- ILLONITE	PYRT	QRTZ
Count	5802	5802	5802	5802	5802	5802	5802	5802	5802	5802
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	0.00	0.53	0.81	0.13	0.00	0.65	0.13	0.71	0.03	0.78
Mean	0.00	0.09	0.03	0.03	0.00	0.35	0.00	0.13	0.01	0.27
Mode	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Median	0.00	0.05	0.00	0.03	0.00	0.37	0.00	0.07	0.01	0.27
Standard Deviation	0.00	0.09	0.09	0.02	0.00	0.11	0.01	0.14	0.00	0.14
25%	0.00	6.50	2.08	2.32	0.00	26.98	0.23	9.59	0.69	20.60
50%	0.00	0.02	0.00	0.01	0.00	0.30	0.00	0.01	0.01	0.17
75%	0.00	0.05	0.00	0.03	0.00	0.37	0.00	0.07	0.01	0.27
Kurtosis	-	3.00	21.79	-0.37	-	0.47	39.13	0.56	0.40	-0.43
Skewness	-	1.76	4.47	0.55	-	-0.70	5.15	1.17	0.57	0.05

Table 8: Statics of the lithology volumes in data set number 2.

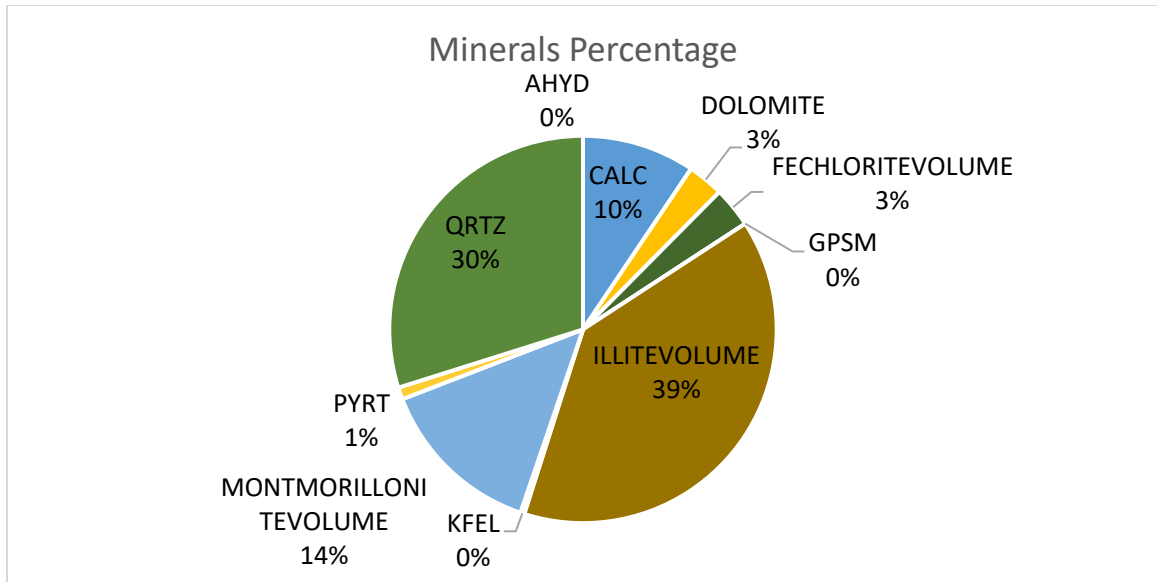


Figure 13: Percentage for the volumes of mineral in data set number 2.

- Compiled Data Set:

As the compiled data set is a combination of both data set number 1 and number 2, it has similar ranges for the variable measures of the two data sets. This can be identified from the statistics number for the compiled data in table-9.

Index	Depth	Gamma Ray	Caliper	Neutron Porosity	Photoelectric	Density	Total NMR Porosity
Count	19735	19735	19735	19735	19735	19735	19735
Minimum	5417.00	14.65	7.89	0.03	2.52	2.43	0.20
Maximum	8900.00	222.36	9.10	0.26	5.41	2.85	14.11
Mean	7244.13	101.92	8.85	0.17	3.25	2.59	6.10
Mode	N/A	101.56	8.83	0.18	3.14	2.63	5.29
Median	7255.50	105.73	8.85	0.18	3.14	2.58	6.09
Standard Deviation	968.33	21.84	0.08	0.03	0.42	0.06	2.23
25%	6433.38	93.22	8.82	0.16	2.99	2.55	4.54
50%	7255.50	105.73	8.85	0.18	3.14	2.58	6.09
75%	8077.88	114.34	8.90	0.19	3.36	2.63	7.62
Kurtosis	-1.13	2.63	26.66	2.25	4.29	1.02	-0.21
Skewness	-0.06	-0.75	-2.88	-1.31	1.90	0.65	0.14

Table 9: Statics of the variables in the compiled data set

a. Hypothesis Testing

One of the steps to support the statistical analysis in machine learning is the hypothesis testing. The objective of the hypothesis testing in machine learning is validating the initial assumption of having a relationship between the input variables and the output variables. The hypothesis testing consists of two components, which are the null hypothesis and the alternative hypothesis [91].

The Null Hypothesis (denoted as H_0) is a statistical assumption suggests that there are no effects or relationships between the input variables (independent variables) and the output variables (dependent variables) within a set of data. This concept denies the existence of any specific data patterns or trends for the analyzed data and that the observed variable correlations are related to coincidental reasons. Generally, the approval of this statement does not align with the machine learning approach and does not support the objective of this study, which is prediction of total NMR porosity from conventional logs. For this reason, a rejection of the null hypothesis is supporting this research. The opposite assumption of the null hypothesis, which indicates that the null hypothesis is rejected, is known as the alternative hypothesis (denoted as H_1 or $\bar{H}$) [23].

An approach to evaluate the hypothesis testing and determine if the null hypothesis is accepted or rejected is calculating the p-value. The p-value is used to measure the probability of an event to occur coincidentally by evaluating a sample of the collected data. It helps assessing the level of significance and effect of the null hypothesis. A common

value for the p-value is 5%, where if it exceeded 5% that means the validity of the null hypothesis [23].

The formula for the null hypothesis and the alternative hypothesis can be seen in the below equations [Eq.7 & Eq. 8]:

$$H_0: p = p_o \quad \text{[Eq. 7]}$$

$$H_1: p > p_o \quad \text{[Eq. 8]}$$

Where:

H_0 : The null hypothesis

H_1 : The alternative hypothesis

p : Calculated p-value

p_o : Selected p-value (=0.05)

In order to apply the hypothesis testing on the relationship between total NMR porosity (output variable) and other log measurements (input variables), the p-value for data samples from the three data sets have been computed using the below python code. In the code, the total porosity was identified as y and the all the other logs were defined as X . The results of the hypothesis testing can be seen in the tables-10 to 12 below.

Code:

```
X = df.drop('T2PTOT',axis=1)
y = df['T2PTOT']
#Checking the Normal Distribution for the thesis data - Shapiro

def checkNormal(p):
    alpha = 0.05
    if p > alpha:
        print("Sample looks normally distributed (fail to reject H0)")
    else:
        print("Sample does not look normally distributed (reject H0)")
```

```

from scipy import stats

w, p = stats.shapiro(X)
checkNormal(p)
w, p = stats.shapiro(y)
checkNormal(p)
w, p = stats.shapiro(df["DEPT"])
checkNormal(p)
w, p = stats.shapiro(df["GR"])
checkNormal(p)
w, p = stats.shapiro(df["CALI"])
checkNormal(p)
w, p = stats.shapiro(df["NPHI"])
checkNormal(p)
w, p = stats.shapiro(df["PE"])
checkNormal(p)
w, p = stats.shapiro(df["RHOB"])
checkNormal(p)

```

- Data Set Number 1:

Input Variable	Hypothesis Testing Result
Depth	Sample does not look normally distributed (reject H0)
Gamma Ray	Sample does not look normally distributed (reject H0)
Caliper	Sample does not look normally distributed (reject H0)
Neutron Porosity	Sample does not look normally distributed (reject H0)
Photoelectric	Sample does not look normally distributed (reject H0)
Density	Sample does not look normally distributed (reject H0)

Table 10: Hypothesis testing results for the variables in data set number 1 with respect to total NMR porosity.

- Data Set Number 2:

Input Variable	Hypothesis Testing Result
Depth	Sample does not look normally distributed (reject H0)
Gamma Ray	Sample does not look normally distributed (reject H0)
Caliper	Sample does not look normally distributed (reject H0)
Neutron Porosity	Sample does not look normally distributed (reject H0)
Photoelectric	Sample does not look normally distributed (reject H0)
Density	Sample does not look normally distributed (reject H0)
Deep Laterolog (LLD)	Sample does not look normally distributed (reject H0)
Shallow Laterolog (LLS)	Sample does not look normally distributed (reject H0)
Mean Spherical focused Log (MSFL)	Sample does not look normally distributed (reject H0)
Potassium (POTA)	Sample does not look normally distributed (reject H0)
Uranium (URAN)	Sample does not look normally distributed (reject H0)
Thorium (THOR)	Sample does not look normally distributed (reject H0)
Compressional wave traveltime (DTC)	Sample does not look normally distributed (reject H0)
Shear wave traveltime (DTS)	Sample does not look normally distributed (reject H0)
AHYD	Sample looks normally distributed (fail to reject H0)
CALC	Sample does not look normally distributed (reject H0)
DOLOMITE	Sample does not look normally distributed (reject H0)
FECHLORITE	Sample does not look normally distributed (reject H0)
GPSM	Sample looks normally distributed (fail to reject H0)
ILLITE	Sample does not look normally distributed (reject H0)
KFEL	Sample does not look normally distributed (reject H0)
MONTMORILLONITE	Sample does not look normally distributed (reject H0)
PYRT	Sample does not look normally distributed (reject H0)
QRTZ	Sample does not look normally distributed (reject H0)

Table 11: Hypothesis testing results for the variables in data set number 2 with respect to total NMR porosity.

- Compiled Data Set:

Input Variable	Hypothesis Testing Result
Depth	Sample does not look normally distributed (reject H0)
Gamma Ray	Sample does not look normally distributed (reject H0)
Caliper	Sample does not look normally distributed (reject H0)
Neutron Porosity	Sample does not look normally distributed (reject H0)
Photoelectric	Sample does not look normally distributed (reject H0)
Density	Sample does not look normally distributed (reject H0)

Table 12: Hypothesis testing results for the variables in the compiled data set with respect to total NMR porosity.

Hypothesis Testing Results Summary:

From the tables, the null hypothesis has been rejected for all the variables in the three data sets except for the anhydrite volume (AHYD) and gypsum volume (GPSM) measurements from data set number 2. The rejection of the null hypothesis indicates to the relationship between the total NMR porosity variable and the other variables in the three data sets. As for anhydrite and gypsum volumes, the acceptance of the null hypothesis could be caused by the zero measurements as showing in table-8 as the tested data is obtained from carbonate reservoirs.

b. Correlation Coefficients Calculations

As the result of the hypothesis testing indicated to the relationships between the input variables from the three data set with the total NMR porosity, the next step is to measure these relationships. One approach to evaluate the degree of a relationship is by calculating the correlation coefficient. The correlation coefficient is a statistical measure that evaluates the strength of the relationship between two variables [62]. There are three types of relationships that are described by the correlation coefficients, which are positive correlation, negative correlation, and non-correlated. The positive correlation indicates to

a relationship between two variables, where the value of one variable increases with the increment of the second variable. On the other hand, the negative relationship indicates to decrease in the value of one variable once the value of the second variable increase. As for the non-correlated relationship, this means that there is no relationship between the two variables. The value of the correlation coefficients varies between 1 and -1, where the total positive correlation equal 1, the total negative correlation equals -1, and the no correlation equals to 0. There are several types of correlation coefficients that can be used to evaluate the relationships between the input variables and the output variables for any machine learning model. In this study, the calculation of the correlation coefficient will be used by determined by using Pearson's correlation coefficient. Pearson's correlation coefficient, which is denoted as r , is measure used to evaluate strength and the direction of the liner relationship between an input variable and an output variable. The equation of Pearson's correlation coefficient is available in [Eq. 9] below [31].

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad [\text{Eq. 9}]$$

Where:

r : Pearson's correlation coefficient

n : Number of data points

X_i : Data value of X data point (input variable)

$\bar{X}$: The mean of X data points

Y_i : Data value of Y data point (output variable)

$\bar{Y}$: The mean of Y data points

The Pearson's correlation coefficient has been computed for the three data sets using the below python code and the calculated correlation coefficient for the variables from the data sets with NMR total porosity is available in the tables (13-15) below.

```

#correlation coefficient calculations with T2PTOT
rho_p_DEPT= pearsonr(df.DEPT, y)
rho_p_GR= pearsonr(df.GR, y)
rho_p_CALI = pearsonr(df.CALI, y)
rho_p_NPHI = pearsonr(df.NPHI, y)
rho_p_PE = pearsonr(df.PE, y)
rho_p_RHOB = pearsonr(df.RHOB, y)
rho_p_LLD= pearsonr(df.LLD, y)
rho_p_LLS= pearsonr(df.LLS, y)
rho_p_MSFL = pearsonr(df.MSFL, y)
rho_p_POTA = pearsonr(df.POTA, y)
rho_p_THOR = pearsonr(df.THOR, y)
rho_p_URAN = pearsonr(df.URAN, y)
rho_p_DTC= pearsonr(df.DTC, y)
rho_p_DTST= pearsonr(df.DTST, y)
rho_p_AHYD = pearsonr(df.AHYD, y)
rho_p_CALC = pearsonr(df.CALC, y)
rho_p_DOLOMITE = pearsonr(df.DOLOMITE, y)
rho_p_FECHLORITEVOLUME = pearsonr(df.FECHLORITEVOLUME, y)
rho_p_GPSM= pearsonr(df.GPSM, y)
rho_p_ILLITEVOLUME= pearsonr(df.ILLITEVOLUME, y)
rho_p_KFEL = pearsonr(df.KFEL, y)
rho_p_MONTMORILLONITEVOLUME = pearsonr(df.MONTMORILLONITEVOLUME, y)
rho_p_PYRT = pearsonr(df.PYRT, y)
rho_p_QRTZ = pearsonr(df.QRTZ, y)
#Print CC for GR with T2PTOT
print ("Pearsons correlation for DEPT: ",rho_p_DEPT)
print ("Pearsons correlation for GR: ",rho_p_GR)
print ("Pearsons correlation for CALI: ",rho_p_CALI)
print ("Pearsons correlation for NPHI: ",rho_p_NPHI)
print ("Pearsons correlation for PE: ",rho_p_PE)
print ("Pearsons correlation for RHOB: ",rho_p_RHOB)
print ("Pearsons correlation for LLD: ",rho_p_LLD)
print ("Pearsons correlation for LLS: ",rho_p_LLS)
print ("Pearsons correlation for MSFL: ",rho_p_MSFL)
print ("Pearsons correlation for POTA: ",rho_p_POTA)
print ("Pearsons correlation for THOR: ",rho_p_THOR)
print ("Pearsons correlation for URAN: ",rho_p_URAN)
print ("Pearsons correlation for DTC: ",rho_p_DTC)
print ("Pearsons correlation for DTST: ",rho_p_DTST)
print ("Pearsons correlation for AHYD: ",rho_p_AHYD)
print ("Pearsons correlation for CALC: ",rho_p_CALC)
print ("Pearsons correlation for DOLOMITE: ",rho_p_DOLOMITE)
print ("Pearsons correlation for FECHLORITEVOLUME:
",rho_p_FECHLORITEVOLUME)

```

```

print ("Pearsons correlation for GPSM: ", rho_p_GPSM)
print ("Pearsons correlation for ILLITEVOLUME: ", rho_p_ILLITEVOLUME)
print ("Pearsons correlation for KFEL: ", rho_p_KFEL)
print ("Pearsons correlation for MONTMORILLONITEVOLUME:
", rho_p_MONTMORILLONITEVOLUME)
print ("Pearsons correlation for PYRT: ", rho_p_PYRT)
print ("Pearsons correlation for QRTZ: ", rho_p_QRTZ)

```

- Data Set Number 1:

Input Variable	Pearson's Correlation Coefficient	Correlation Type
Depth	-0.433	Negative Correlation
Gamma Ray	0.105	Positive Correlation
Caliper	0.147	Positive Correlation
Neutron Porosity	0.282	Positive Correlation
Photoelectric	-0.299	Negative Correlation
Density	-0.382	Negative Correlation

Table 13: Pearson's correlation coefficient calculations for the variables in data set number 1 with respect to total NMR porosity.

- Data Set Number 2:

Input Variable	Pearson's Correlation Coefficient	Correlation Type
Depth	-0.380	Negative Correlation
Gamma Ray	0.140	Positive Correlation
Caliper	0.046	Positive Correlation
Neutron Porosity	0.296	Positive Correlation
Photoelectric	-0.253	Negative Correlation
Density	-0.369	Negative Correlation
Deep Laterolog (LLD)	-0.050	Negative Correlation
Shallow Laterolog (LLS)	-0.055	Negative Correlation
Mean Spherical focused Log (MSFL)	-0.025	Negative Correlation
Potassium (POTA)	-0.001	Negative Correlation
Thorium (THOR)	-0.00004	Negative Correlation
Uranium (URAN)	0.209	Positive Correlation

Compressional wave traveltime (DTC)	0.359	Positive Correlation
Shear wave traveltime (DTS)	0.442	Positive Correlation
AHYD	nan	Non-Correlated
CALC	-0.109	Negative Correlation
DOLOMITE	-0.187	Negative Correlation
FECHLORITE	-0.231	Negative Correlation
GPSM	nan	Non-Correlated
ILLITE	0.022	Positive Correlation
KFEL	-0.071	Negative Correlation
MONTMORILLONITE	0.157	Positive Correlation
PYRT	-0.098	Negative Correlation
QRTZ	0.017	Positive Correlation

Table 14: Pearson's correlation coefficient calculations for the variables in data set number 2 with respect to total NMR porosity.

- Compiled Data Set:

Input Variable	Pearson's Correlation Coefficient	Correlation Type
Depth	-0.421	Negative Correlation
Gamma Ray	0.113	Positive Correlation
Caliper	0.120	Positive Correlation
Neutron Porosity	0.286	Positive Correlation
Photoelectric	-0.286	Negative Correlation
Density	-0.378	Negative Correlation

Table 15: Pearson's correlation coefficient calculations for the variables in the compiled data set with respect to total NMR porosity.

Pearson's Correlation Coefficient Results:

From the previous tables and the below charts, it can be observed that the similar data measurement from the three groups have the same correlation types, where depth, photoelectric, and density measurements have negative correlation and gamma ray, caliper, and neutron porosity measurements have positive correlation with respect to the total NMR porosity. As for the additional log measurements from data set number 2, the correlation coefficients calculations vary between the different types of correlations, where Uranium

(URAN), compressional wave traveltime (DTC), shear wave traveltime (DTS) have positive correlations and other logs have very low negative correlations. As for the lithology log measurements in data set number 2, there are three measurements that have provided positive correlations, which are illite, montmorillonite, and quartz. The remaining measurements have provided negative correlation coefficients except for anhydrite volume (AHYD) and gypsum volume (GPSM) as they are non-correlated with NMR porosity due to the zero values.

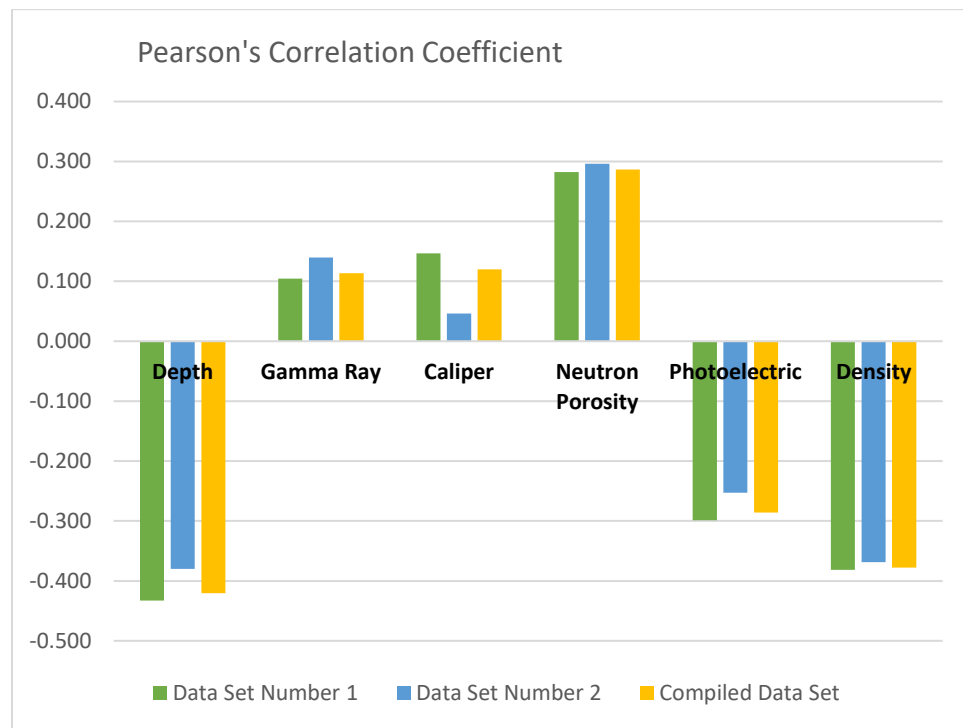


Figure 14: Comparison of Pearson's correlation coefficient for the variables from the three data sets with respect to total NMR porosity.

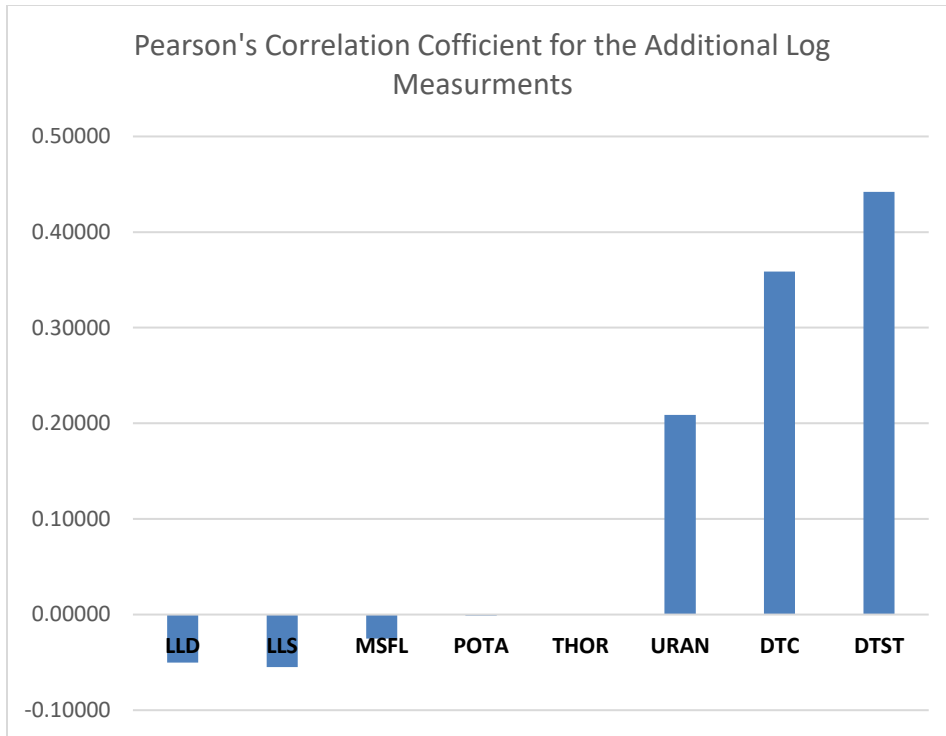


Figure 15: Pearson's correlation coefficient for the variables the additional log measurements from data set number 2.

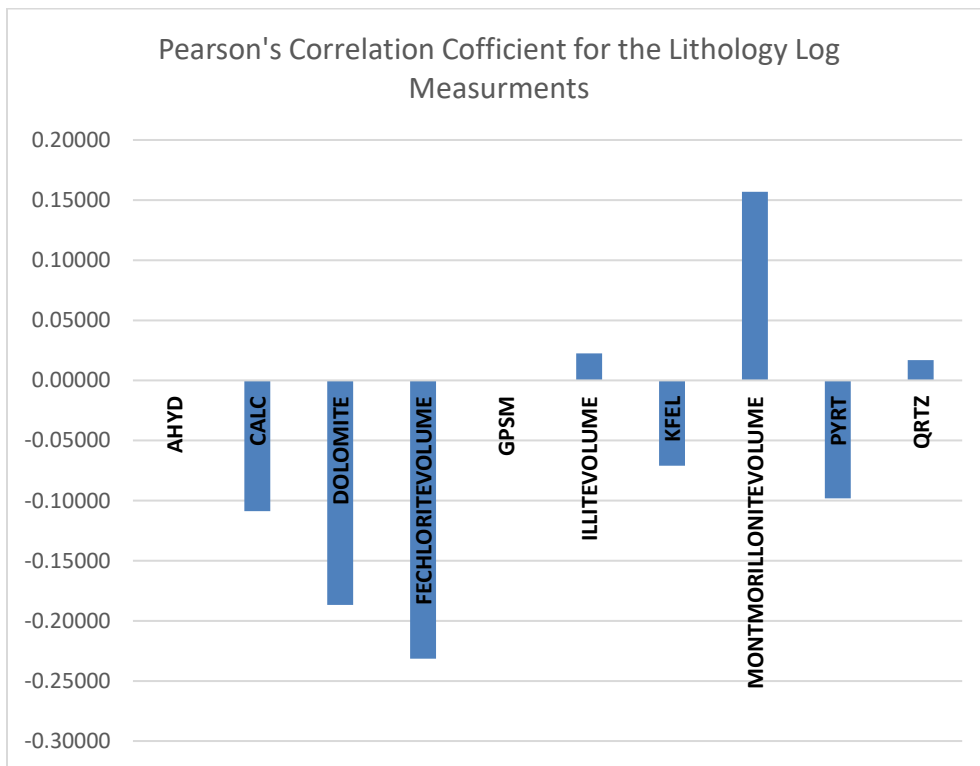


Figure 16: Pearson's correlation coefficient for the variables the lithology log measurements from data set number 2.

c. Data Visualization

Data visualization can provide additional information during the data exploration. It enhances the ability to understand the data by availing graphical meaning for the data. This will expediate the process of allocating the different patterns, trends, and outliers within the data. In addition, it will help to identify the complex relationships between the data variables. Besides that, several plots can assist in taking analysis decisions by providing an overview for the data distributions [105]. In this section, some of the data plots and their obtained information.

- Data Set Number 1:

A pair plot chart for the data set number 1 is displayed in figure-15. The previous results of the relationships between the different variables in the data set can be confirmed from the pair plot chart. The initial observations for the pair plot indicate to a general positive correlation between total porosity with gamma ray and neutron porosity logs. On the other hand, a general negative correlation between total porosity with photoelectric and bulk density logs can be noticed. In addition, the heatmap in figure-16 confirms the calculations of the Pearson's correlation coefficient. Generally, the histogram plots in figure-17 display normal distributions for the total NMR porosity in data set number 1. This eliminates the necessity to identify the outliers as the data distributions behave normally.

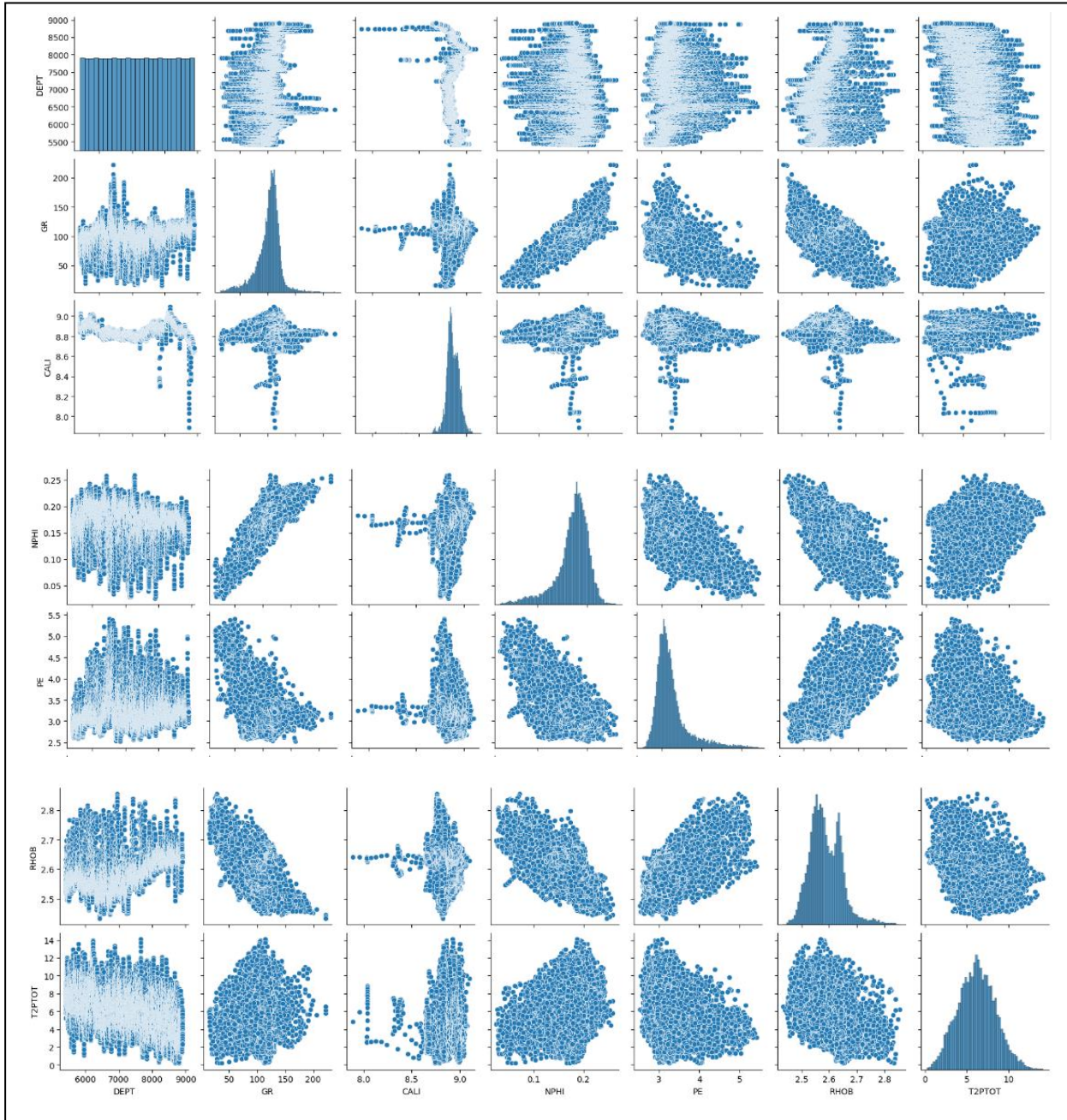


Figure 17: Pair plot for the correlations between the variables in data set number 1.

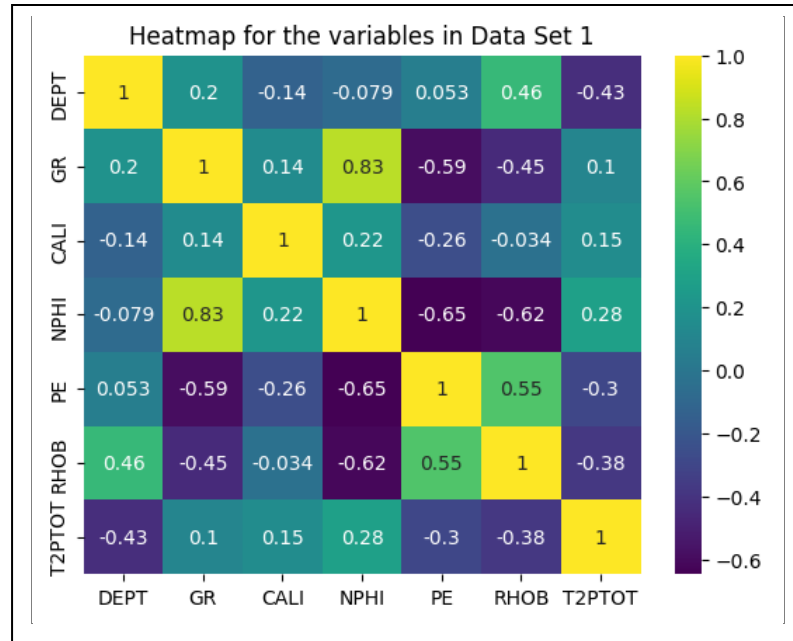


Figure 18: Heatmap for the correlations between the variables in data set number 1.

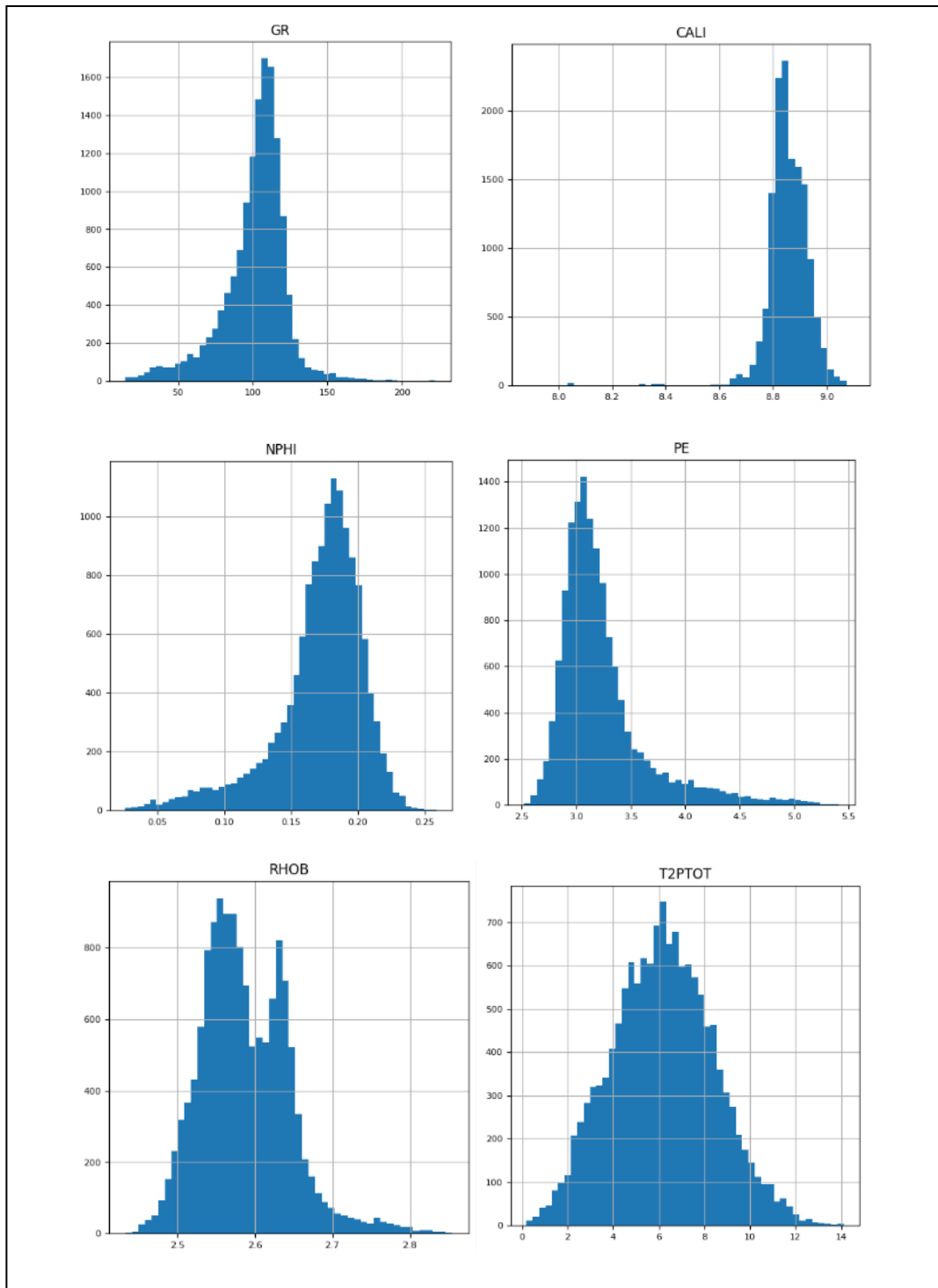


Figure 19: Histogram for data distribution of the logs' variables in data set number 1.

(Note: Depth was not included as it is the index for the values of the other variables)

- Data Set Number 2:

Figures-18 to figure-20 demonstrate the relationships between the variables in data set number 2 in pair plots charts. Figure-18 confirms the positive correlations between gamma ray and the neutron porosity logs measurements with the total NMR porosity. The same conditions apply for the photoelectric and bulk density log measurements, where they have a negative correlation with total NMR porosity, similar to data set number 1. As for the additional logs and the lithology measurements in data set number 2, their correlations with NMR porosity are not clear from the pair plots in figure-19 and figure-20 except for compressional wave traveltime (DTC) and shear wave traveltime (DTST). These two measurements are confirming the Pearson's correlation coefficient calculations in table-14. The correlation coefficients calculations for all the variables in data set number 2 can be confirmed as well from the heatmap in figure-21.

Figure-22 to figure-24 display the distributions for the number of data points for the variables in data set number 2. The common logs between data set number 1 and data set number 2 are providing the same data distribution, where the histogram plot shows normal distribution for gamma ray and total NMR porosity. In addition, it is displaying a positively skewed distribution for the photoelectric measurements and a negatively skewed distribution for the neutron porosity measurements. In case of the additional log measurements in data set number 2, there is a variation in the data distribution. The measurements of Potassium (POTA) and Shear wave traveltime (DTST) are providing normal data distribution. As for Thorium (THOR) and Compressional wave traveltime (DTC) measurements, they are displaying negatively skewed data distribution. The

positively skewed data distribution is showing for the Uranium (URAN) measurements. Additionally, the distribution for the lithology measurements have been plotted in figure-25. The plot displays positively skewed distribution for Calcite Volume (CALC), Chlorite Volume (FECHLORITEVOLUM), Montmorillonite Volume, Pyrite Volume (PYRT), and Quartz Volume (QRTZ) and negatively skewed distribution for Illite Volume (Illitevolume). The remaining lithology measurements are mostly showing zero values for the logs' readings. This reduce the dependency on these logs within the machine learning model. The overall data distributions do not require any removal for outliers.

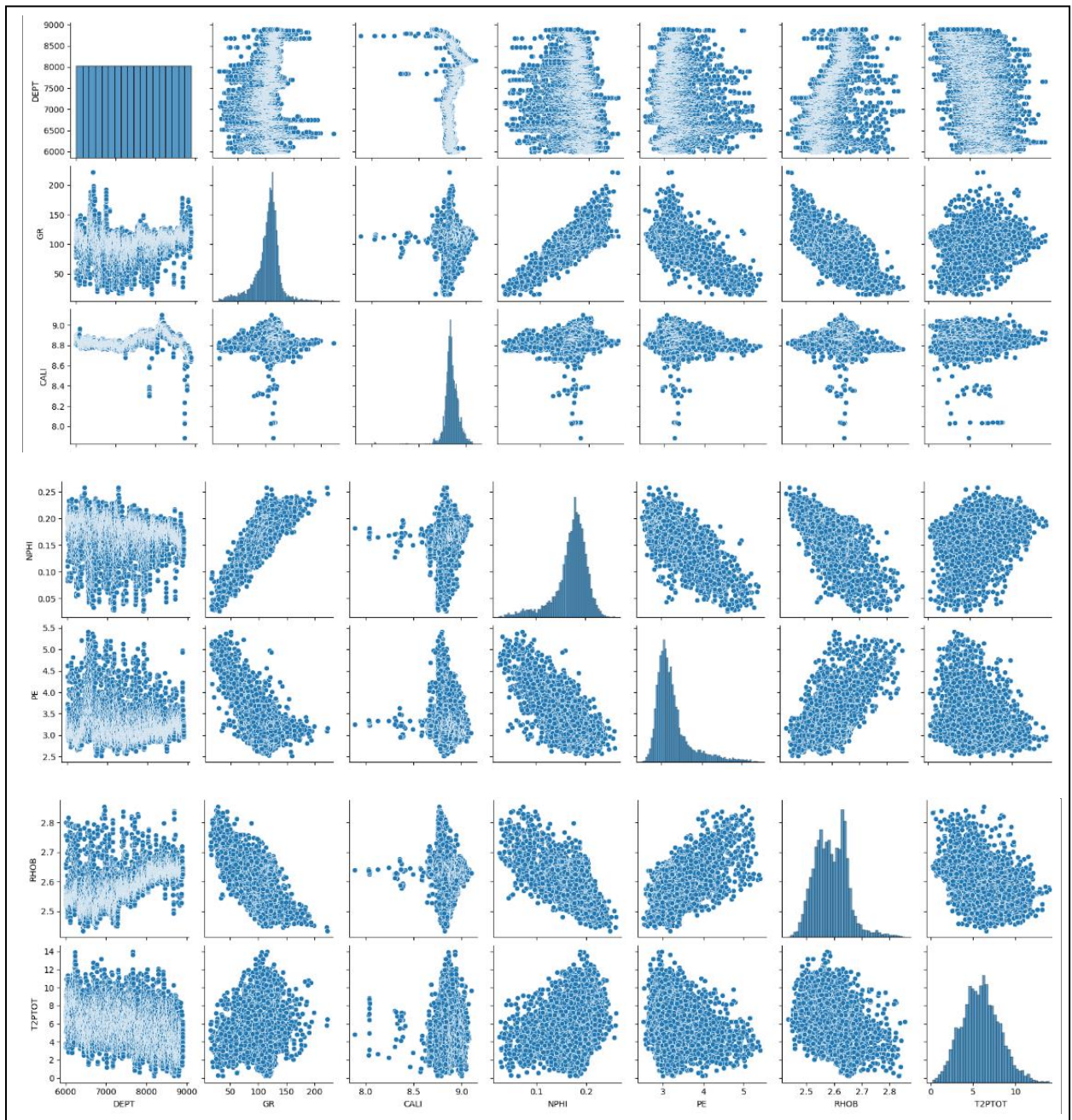


Figure 20: Pair plot for correlations between the logs' variables in data set number 2.

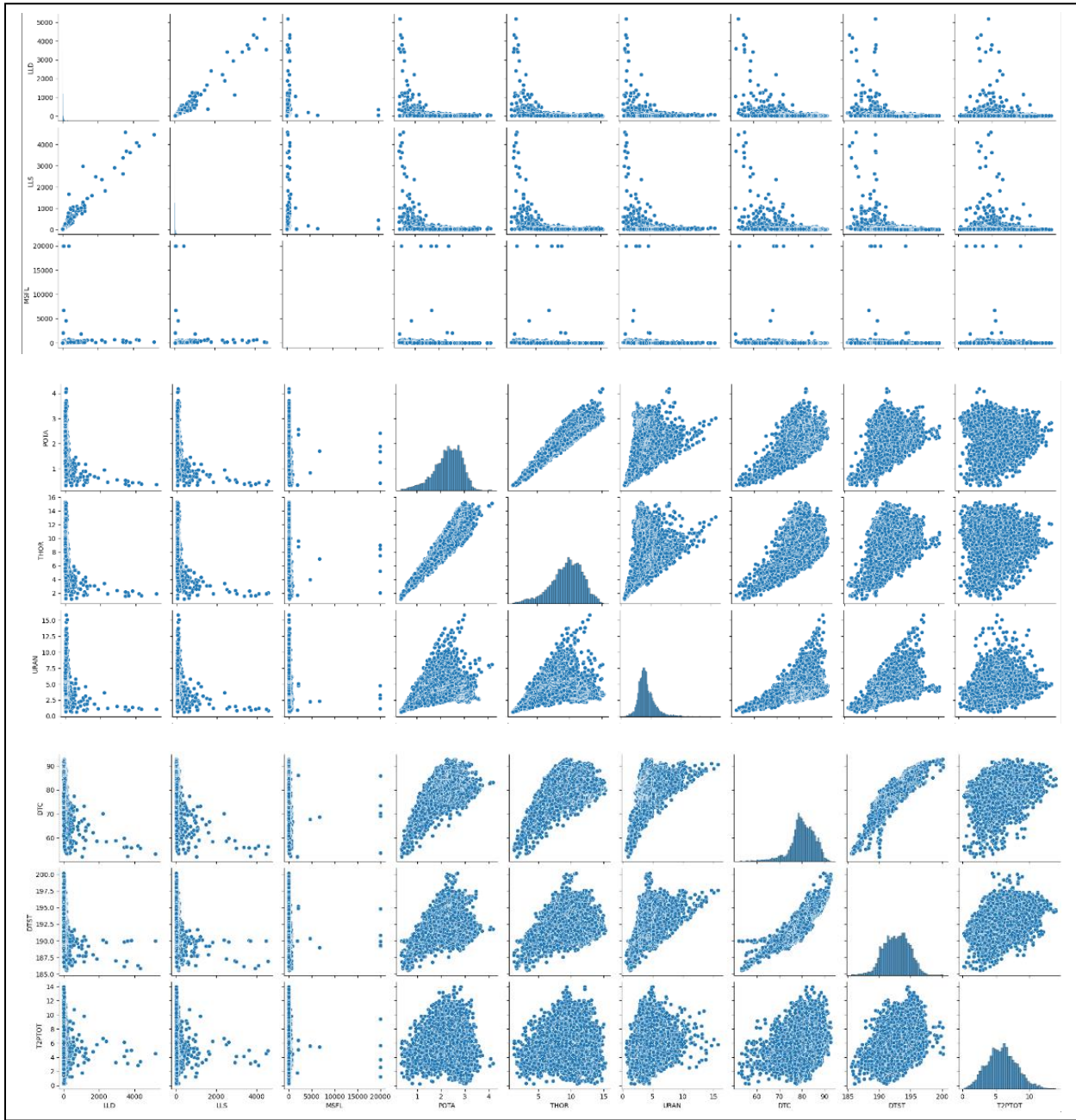


Figure 21: Pair plot for correlations between the additional variables in set number 2.

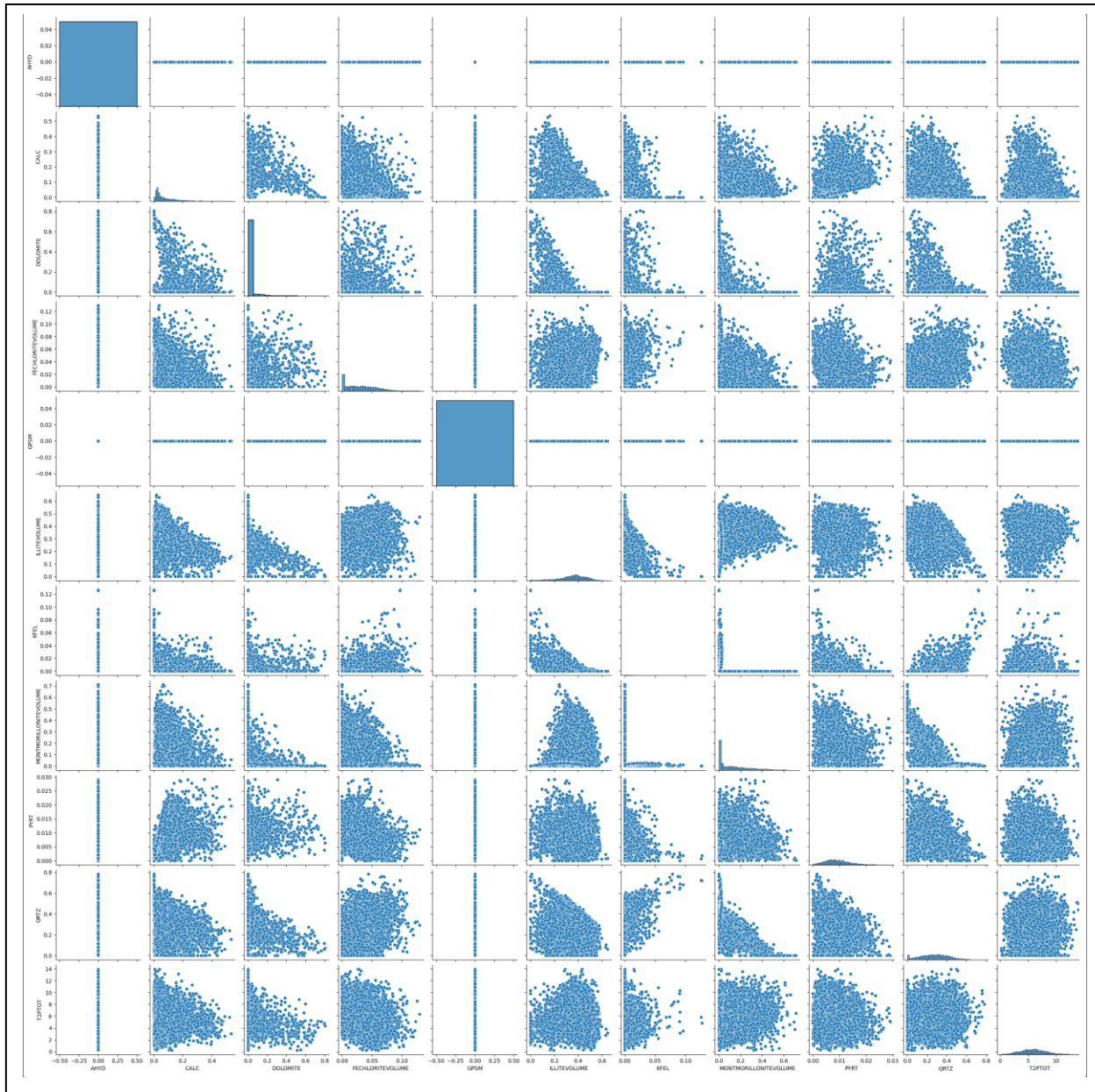


Figure 22: Pair plot for correlations between the lithology volumes in set number 2.

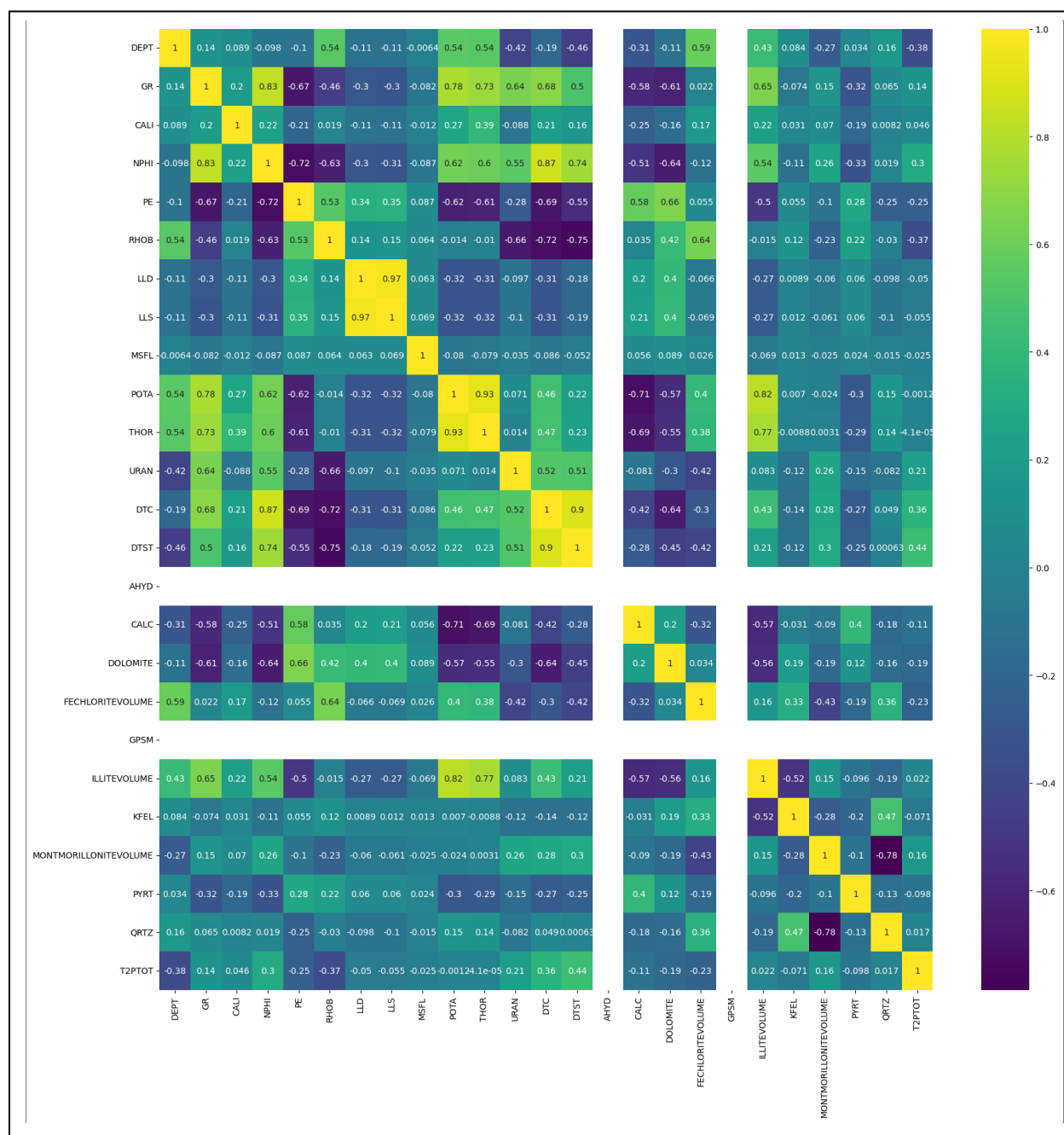


Figure 23: Heatmap for the correlations between the variables in data set number 2.

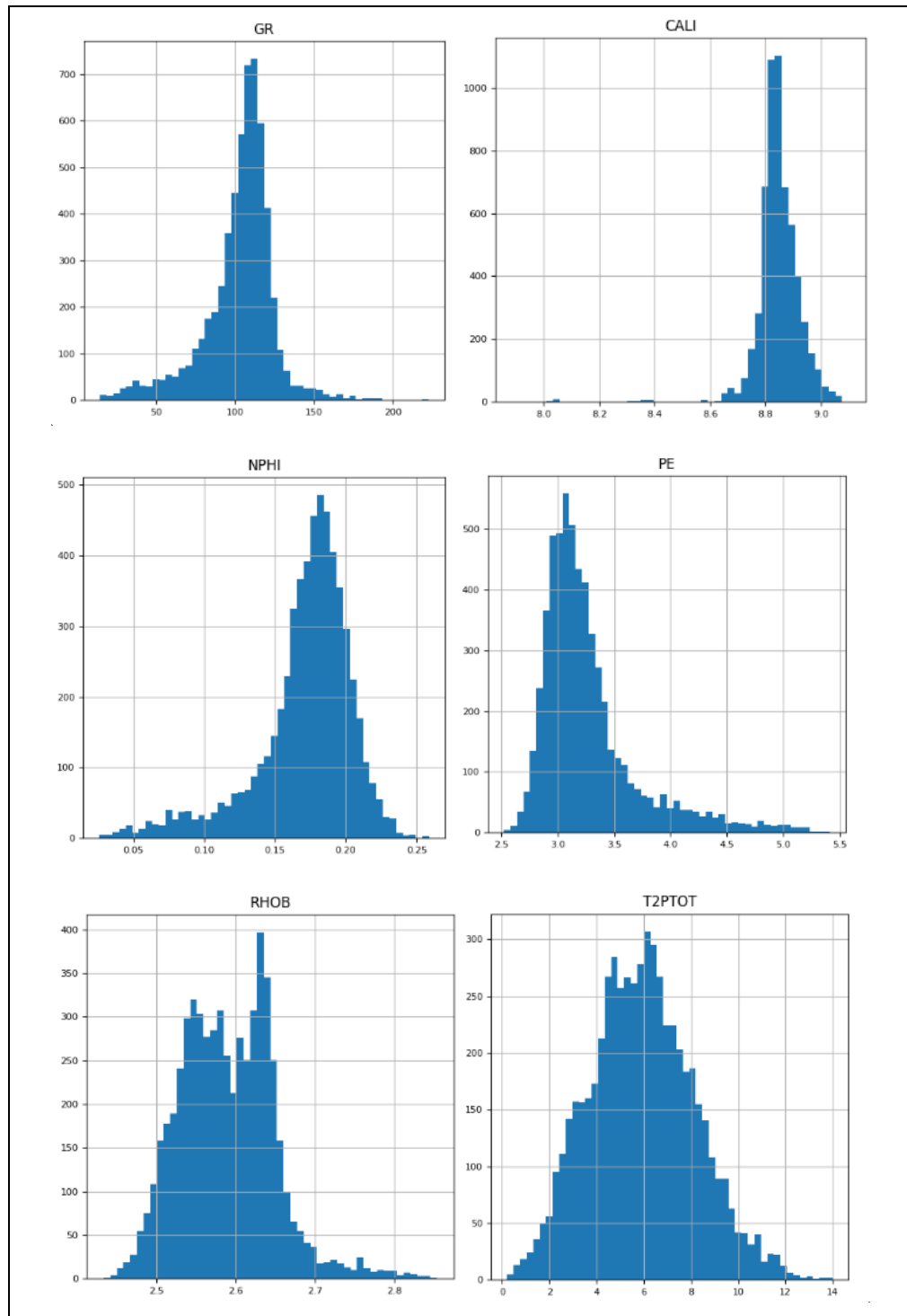


Figure 24: Histogram for the data distribution of the log variables in data set number 2.

(Note: Depth was not included as it is the index for the values of the other variables)

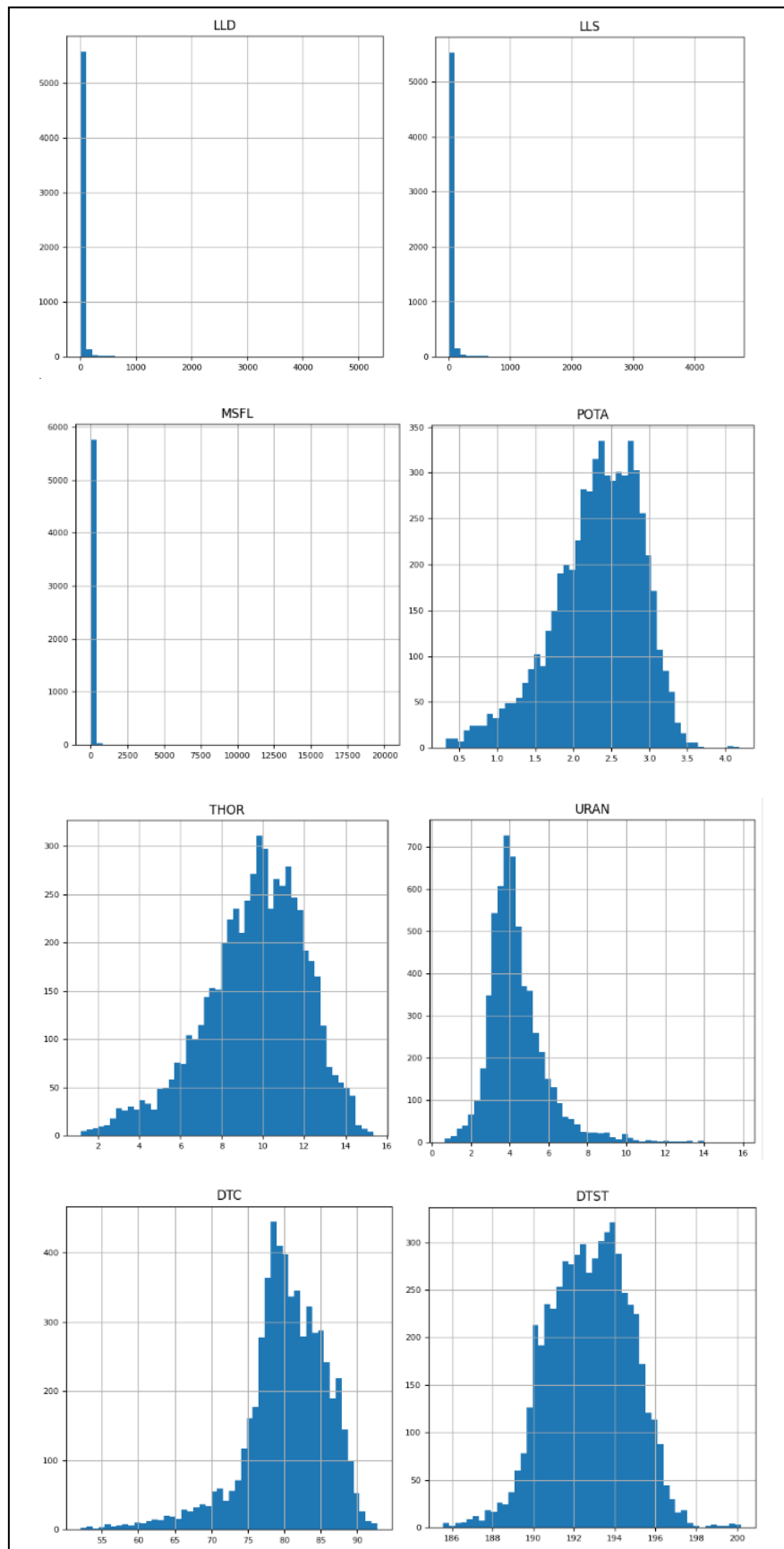


Figure 25: Histogram for data distribution of additional logs' variables in set number 2.

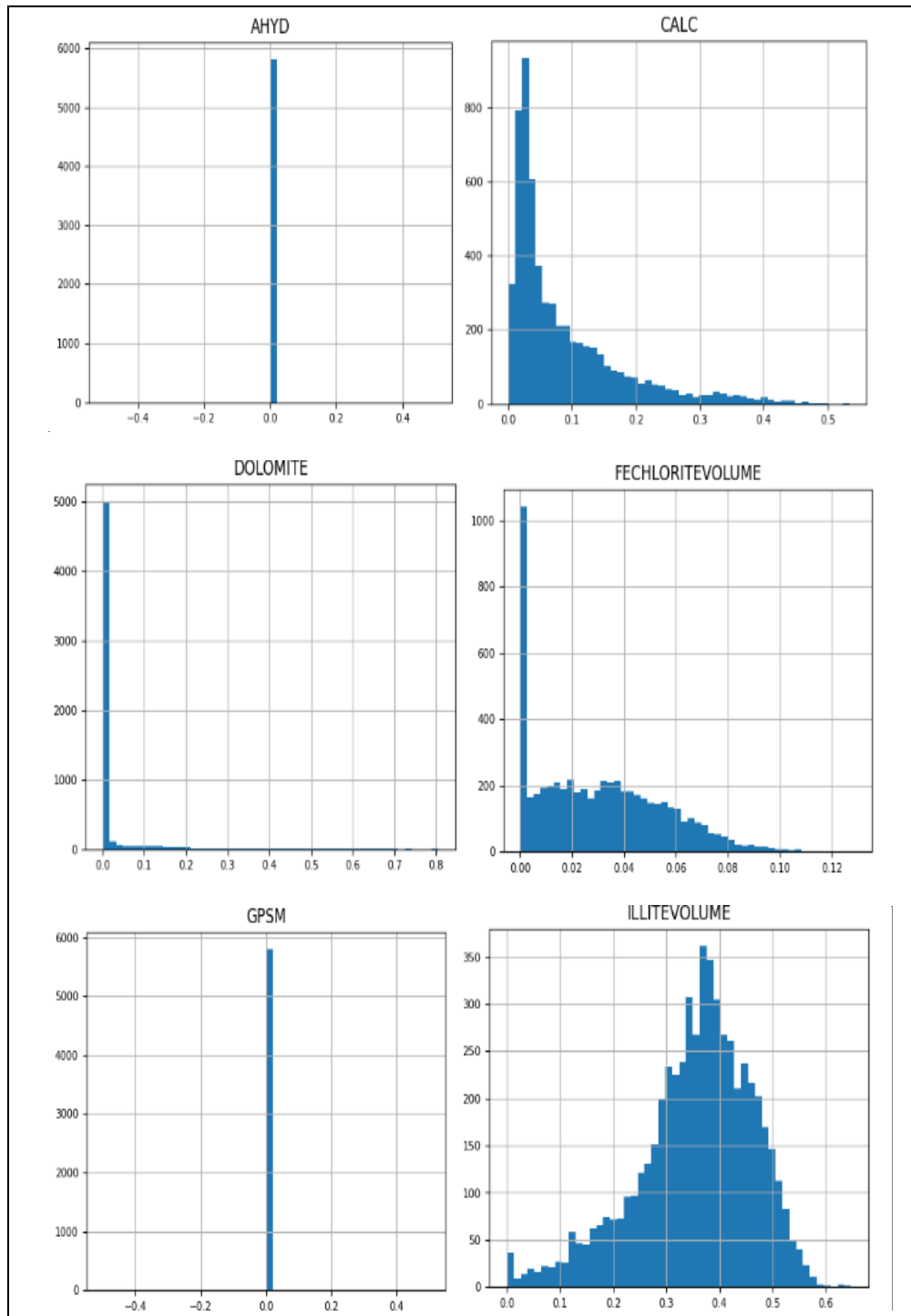


Figure 26: Histogram for data distribution of lithology volumes in set number 2.

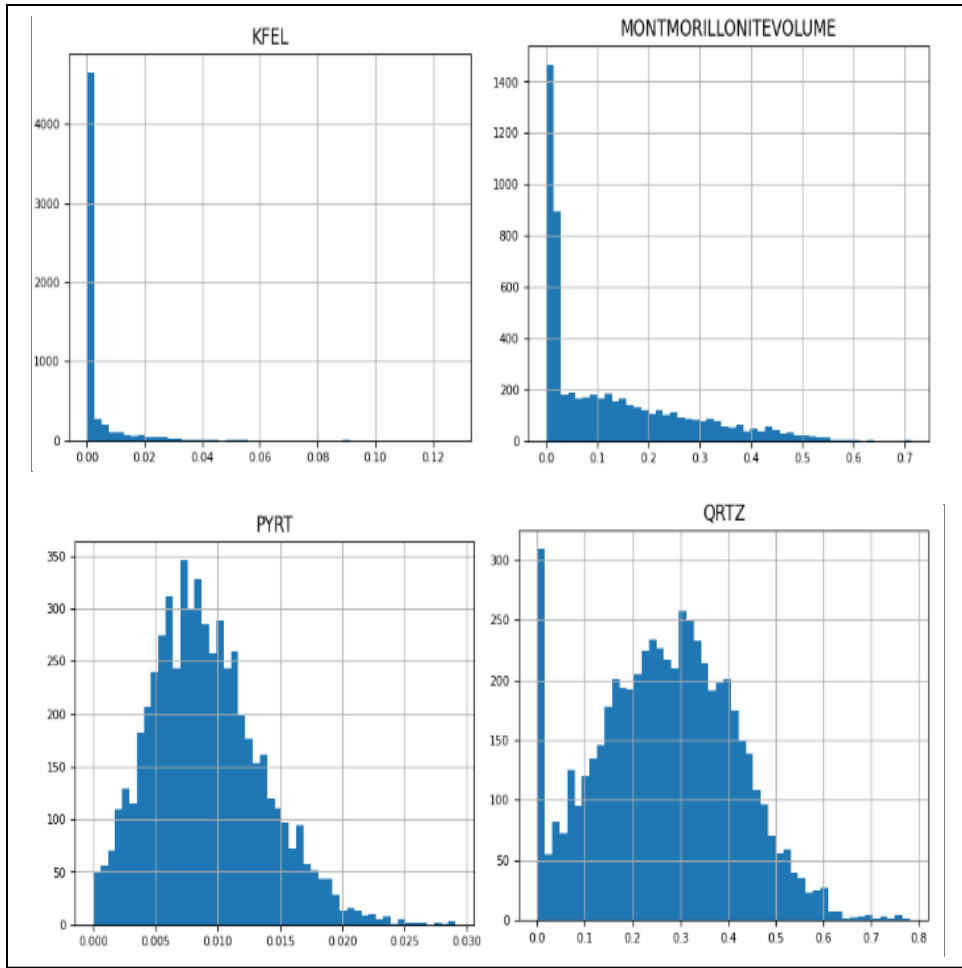


Figure 27: Histogram for data distribution of lithology volumes in set number 2.

- Compiled Data Set:

As the compiled data set is a combination of data set number 1 and data set number 2, the same positive and negative correlations can be seen from the pair plot for the compiled data set in figure-26. The figure supports the Pearson's correlation calculations in table-15, which has been re-confirmed by the heatmap of the correlation coefficients in figure-27. Similarly, the data distribution for variables the compiled data sets is matching the data distribution for the common variables in the two data sets. This can be observed in figure-28. So, the compiled data set does not require any outliers' identification as well.

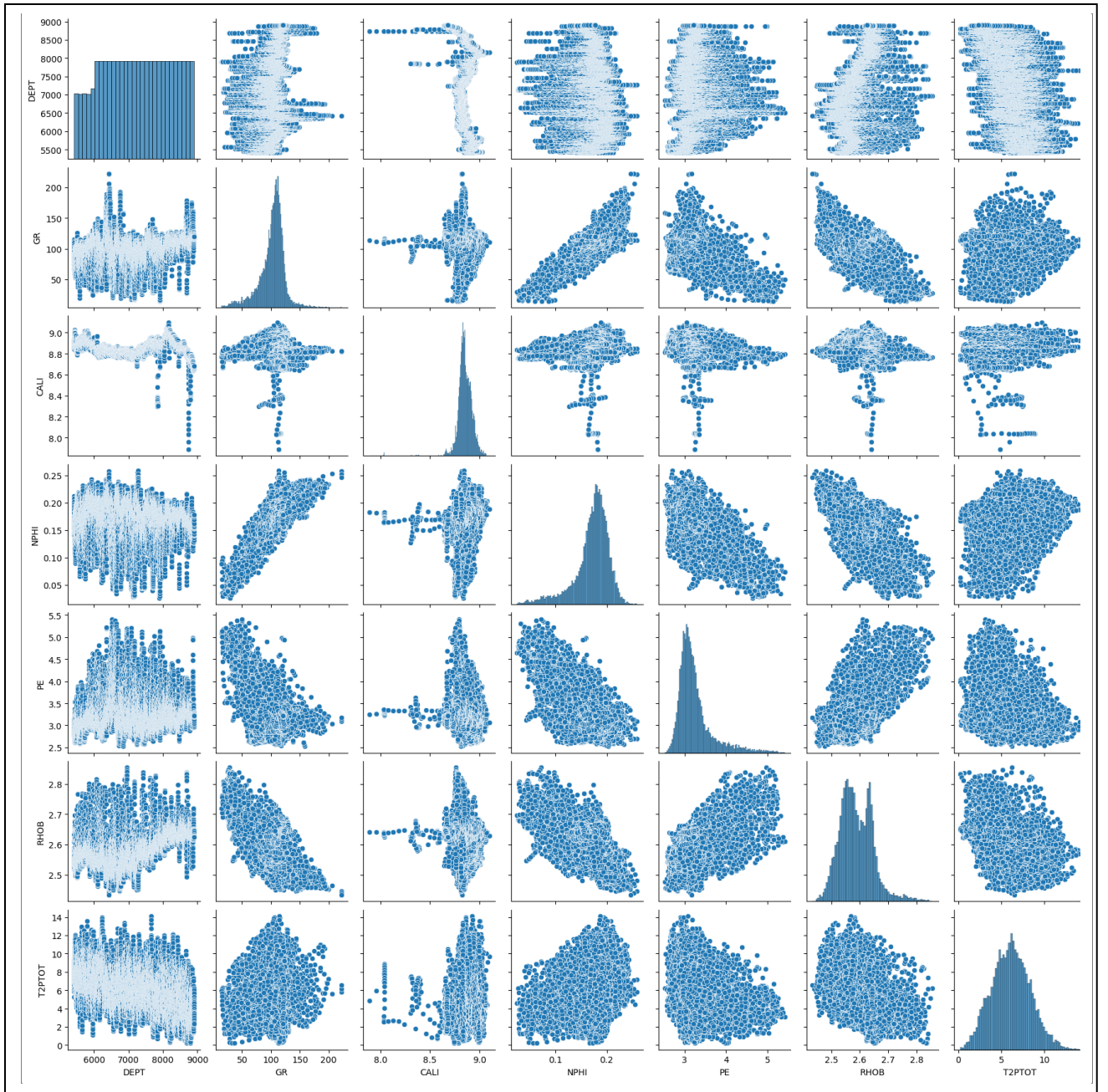


Figure 28: Pair plot for correlations between logs' variables in the compiled set.

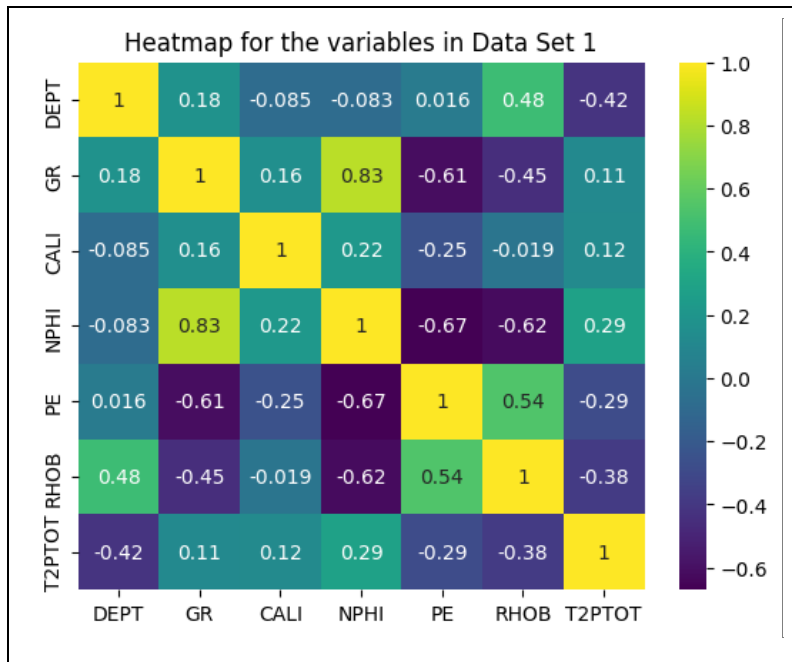


Figure 29: Heatmap for the correlations between the variables in the compiled data set.

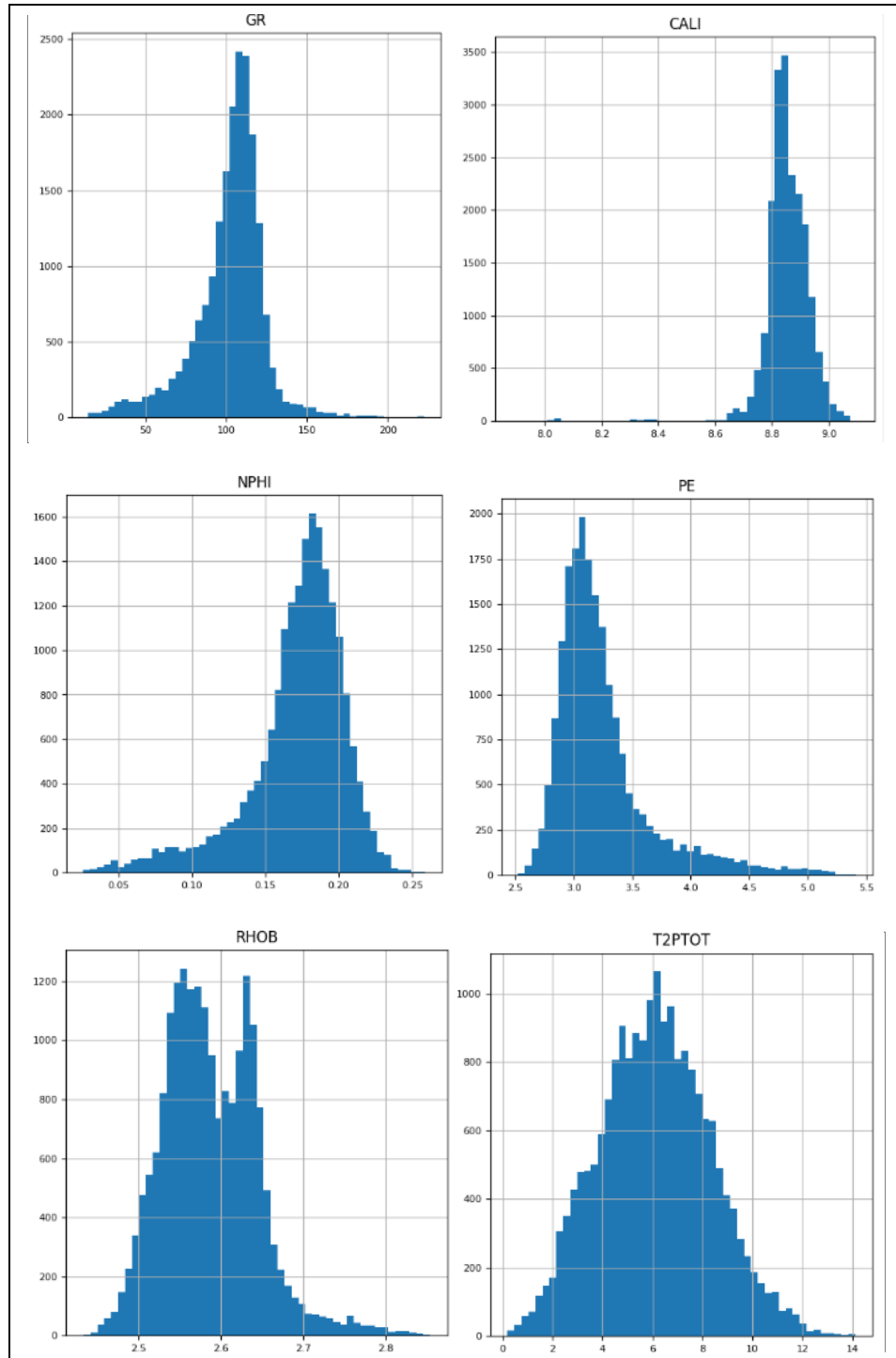


Figure 30: Histogram for data distribution of logs' variables in the compiled set.

(Note: Depth was not included as it is the index for the values of the other variables)

Exploratory Data Analysis Outcomes:

- Lithology measurements from data set 2 established the rock type as carbonate.
- Input variables (data features) and the output variable (total NMR porosity) demonstrated values within acceptable ranges for carbonate reservoirs, ensuring data representativeness.
- A minor portion of data points (less than 0.8%) were removed from Data set number 1 due to inaccurate gamma ray readings.
- The null hypothesis was rejected, confirming the statistical relationship between the output variable (total NMR porosity) and the input variables across all data sets.
- Pearson's Correlation Coefficient validated assumptions regarding both negative and positive correlations between input and output variables, providing a quantitative assessment of these relationships.
- Anhydrite volume (AHYD) and gypsum volume (GPSM) from Data set number 2 exhibited zero values across all data points, confirming them non-correlated with NMR porosity.
- Data visualization techniques visually validated the statistical findings, effectively demonstrating relationships and correlations between input and output variables.
- The output variable (total NMR porosity) exhibited a normal distribution in all three data sets, which confirms the accepted data range for machine learning model.

CHAPTER 5

RESULTS

5.1 Error Metrics

There are several metrics that can be used to evaluate performance of machine learning models. The selection of the error metrics depends on the research problem and the utilized solution. In this study, the metrics that were used to assess the model performance are the correlation coefficient (R^2), absolute average percentage error (AAPE), and root mean square error (RMSE). The correlation coefficient is measuring the accuracy between the actual and the predicted values from the ANN model [Eq.10]. As for the AAPE and RMSE, they are used to calculate the predicted values and estimate the consistency of the ANN model [Eq.11/Eq.12][88/89].

$$R^2 = \frac{n (\sum xy) - (\sum x)(\sum y)}{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]} \quad [\text{Eq. 10}]$$

$$AAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{x - y}{x} \right| \quad [\text{Eq. 11}]$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x - y)^2}{n}} \quad [\text{Eq. 12}]$$

Where:

n : Number of data points

x : Actual values

y : Predicted values

$\sum x$: Sum of all actual values

$\sum y$: Sum of all predicted values

$\sum xy$: Sum of all actual and predicted values

To ensure that the model is evaluated correctly by the error metrics, each data set of three data sets have been divided within each model into two groups, which are training and

testing data. This will allow the model to be trained with training group of data and to be evaluated with the testing group of data. The testing group of data will not be included in the training of the model and they will be considered as new set of data to the model. There are common ratios for the splitting of the data sets into training and testing groups such as 70-30%, 80-20%, 90-10% for training and testing, respectively. One important condition in the splitting of the data sets to the training and testing groups is to conduct the splitting randomly to guarantee that the groups are representing the overall distribution of the overall data set.

Although the evaluation of the model performance will be conducted separately on the training and testing groups, the results from both groups will be utilized to evaluate the overall performance of the model. This step is necessary to determine the direction of results improvements. A general rule to identify the causes of a machine learning model using the resulted errors for training and testing groups is displayed in figure-29. The rule indicates that the case of high errors in both training and testing groups are requiring improvement in the model complexity or adding more input variables (more features). On the other hand, the case of low error for the training group and high error for the testing group indicates a situation of overfitting, which can be resolved by increasing the number of the data points or performing model regularization.

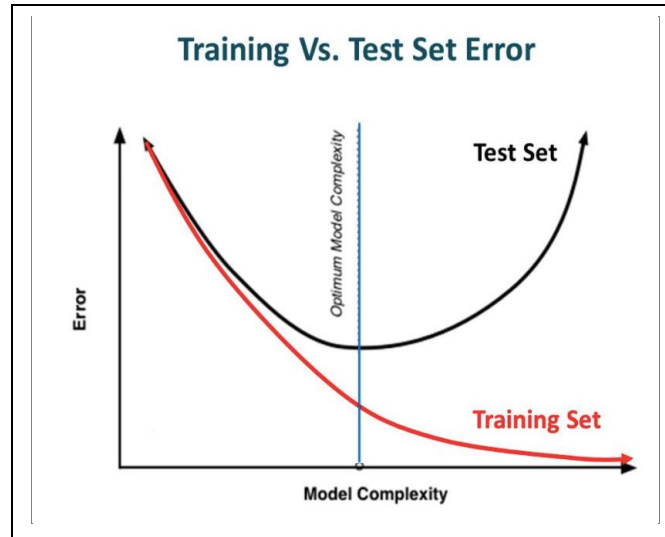


Figure 31: Error evaluation vs model complexity of training and testing groups [60].

5.2 Silhouette Coefficient

Part of developing the machine learning model is separating the data into clusters using the clustering algorithm. In order to evaluate the clustering process, Silhouette Coefficient will be used as the measure to order the results of the clustering process for the different clustering algorithms. Silhouette Coefficient assess the suitability of the data points in regards to the assigned cluster. For data point i , the Silhouette score (s_i) depends on two values, which are the cohesion (a) and the separation (b). The cohesion (a) is the calculation of the average distance of the selected data point i from all the other data points within the same cluster. The separation (b) is the minimum of the calculations for the average distance of the data point i to the data points in other clusters. The calculation of Silhouette score can be completed using [Eq.13]. The score of the equation will ranges from -1 to 1. As the

score values increasing toward 1, this means a well clustered data. On the hand, the values are closer to -1 indicate to the low quality of the clustering process [93].

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad [\text{Eq. 13}]$$

5.3 Results

5.3.1 Primary Models (Model-A)

- Data Set Number 1:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	28%	1.5175	1.9002	27%	1.5170	1.9002
DecisionTreeRegressor	100%	0.0000	0.0000	38%	1.1873	1.7530
DecisionTreeRegressor	100%	2.90E-16	1.53E-15	35%	1.2222	1.7953
RandomForestRegressor	95%	0.3670	0.4846	68%	0.9680	1.2704
XGBRFRegressor	39%	1.3929	1.7478	35%	1.4294	1.7979
XGBRegressor	81%	0.7403	0.9738	55%	1.1650	1.4959
GradientBoostingRegressor	41%	1.3669	1.7139	38%	1.4006	1.7610
LGBMRegressor	58%	1.1469	1.4439	47%	1.2879	1.6302
MLPRegressor	35%	1.4470	1.8066	33%	1.4633	1.8305
SVR	34%	1.4322	1.8229	32%	1.4524	1.8399
keras	28%	1.5112	1.8970	28%	1.5081	1.8951

Table 16: The results of the machine learning models of data set number 1.

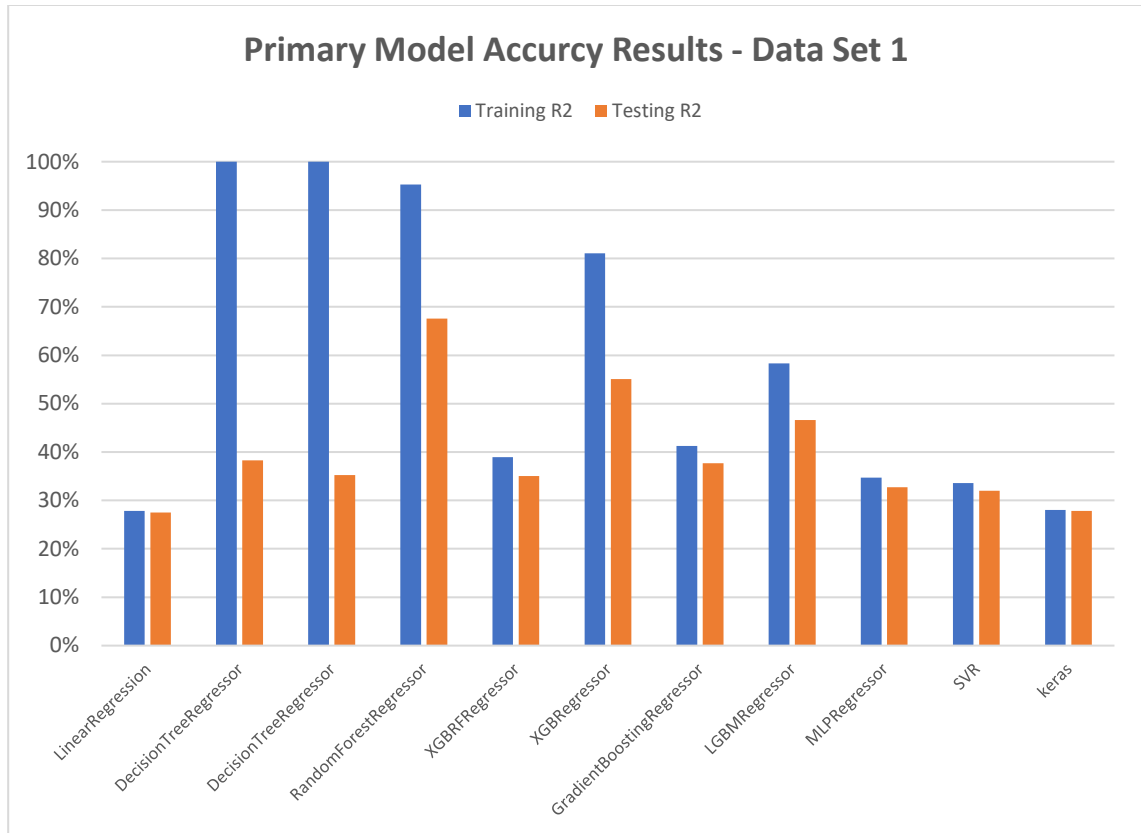


Figure 32: Accuracy comparison of training and testing results for the primary machine learning models of data set number 1.

Results Remarks:

- Decision Tree Regressor models (2 models) have provided the highest values for the correlation coefficient (R^2) of the training group (100%). The given results of (100%) for the training group might indicate an overfitting of the models for the training group. This interpretation is supported by the low values of the correlation coefficient (35%) and the high error values for both absolute average percentage error (AAPE), and root mean square error (RMSE) of the testing group.
- Random Forest Regressor model has provided a good result for the correlation coefficient (R^2) of the training group (95%) and the best results for the correlation coefficient (R^2) of the testing group (68%). In addition, it provided low values of

both absolute average percentage error (AAPE), and root mean square error (RMSE).

- Figure-30 displays variation for the machine learning models accuracy. Some of the models are showing high accuracy for the training group and low accuracy for the testing group, which can be interrupted as overfitting. On the other hand, there are models showing low accuracy for both training and testing groups. This is a sign of low model performance.
- By comparing the accuracy and the error results of the primary machine learning models of data set number 1 with figure-29, further improvements are required to enhance the models results. The improvements include solutions for the overfitting and low performance issues to enhance the machine learning models prediction results.

- Data Set Number 2:

As data set number 2 has more input variables than the other data sets, the following approaches were used in testing data set number 2:

a. Eliminating zero-reading variables:

From the exploratory analysis, it was confirmed that both Anhydrite volume (AHYD) and gypsum volume (GPSM) are non-correlated with NMR porosity. For this reason, both variables were removed from data set number 2 that was tested in the machine learning models.

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	27%	1.5164	1.8982	27%	1.5263	1.8971
DecisionTreeRegressor	100%	0.0000	0.0000	-11%	1.7518	2.3329
DecisionTreeRegressor	100%	3.03E-16	1.40E-15	5%	1.6463	2.1584
RandomForestRegressor	93%	0.4670	0.5966	50%	1.2540	1.5701
XGBRFRegressor	42%	1.3428	1.6815	33%	1.4641	1.8116
XGBRegressor	96%	0.3046	0.4180	45%	1.2942	1.6364
GradientBoostingRegressor	43%	1.3328	1.6688	34%	1.4495	1.7930
LGBMRegressor	76%	0.8541	1.0800	42%	1.3437	1.6794
MLPRegressor	43%	1.3255	1.6732	34%	1.4474	1.8008
SVR	34%	1.3852	1.7965	31%	1.4598	1.8375
keras	-31%	1.9692	2.5390	-33%	1.9693	2.5513

Table 17: The results of the machine learning models of data set number 2 (without the variables of zero readings).

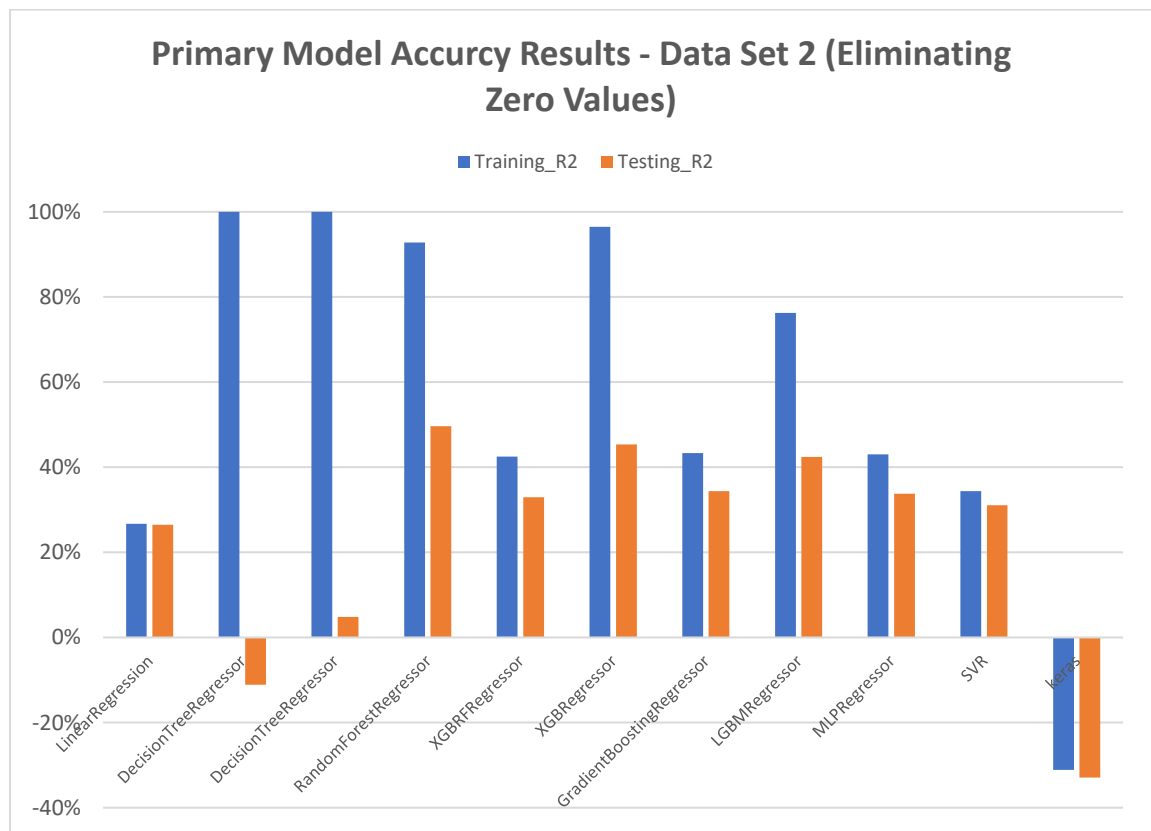


Figure 33: Accuracy comparison of training and testing results for primary machine learning models of data set number 2 (without variables of zero readings).

Results Remarks:

- Similar to data set number 1, the Decision Tree Regressor models have provided high values for the training group (100%) and low values testing group (-11% & 5%) of the correlation coefficient (R^2). Similarly, the resulted values both absolute average percentage error (AAPE), and root mean square error (RMSE) of the testing group are high comparing to the training group. This is an indication of overfitting results.
- The three models Random Forest Regressor, XGBRegressor, and LGBMRegressor have provided good values of the correlation coefficient (R^2) 93%, 96%, and 76% for the training group, respectively. Similarly, the three models have provided better values of the correlation coefficient (R^2) for the testing group than the other machine learning models. These values are not sufficient results for the models, but they are an indication of the models' primary capabilities.
- The keras Artificial Neural Network (ANN) model has provided negative values of the correlation coefficient (R^2) for both training and testing groups. This refers to the negative matching between the actual values and the predicted values for the total NMR porosity.
- There is variation between the machine learning models results with data set number 2. Some of the models are presenting overfitting results, where the training group are providing high accuracy and the testing group are providing low accuracy. Other models are demonstrating low performance of the models as both testing and training groups are delivering low values of the correlation coefficient (R^2).

- The comparisons of the models' error values with figure-29 illustrates that the models' complexity needs to be improved in order to reduce the error values of the training and the testing groups and overcome the low performance issue. In addition, more regularization or increasing the number of data can resolve the overfitting of the results.

b. Eliminating lithology variables:

The main objective of objective of this research is to predict NMR porosity using machine learning models as NMR porosity is lithology-independent. This means that lithology variables should not be utilized as part of testing the machine learning models.

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	26%	1.5288	1.9102	26%	1.5350	1.8997
DecisionTreeRegressor	100%	0.0000	0.0000	2%	1.6415	2.1945
DecisionTreeRegressor	100%	3.03E-16	1.40E-15	2%	1.6501	2.1949
RandomForestRegressor	93%	0.4662	0.5955	50%	1.2444	1.5606
XGBRFRegressor	41%	1.3557	1.7019	34%	1.4606	1.7986
XGBRegressor	94%	0.3904	0.5404	45%	1.3125	1.6410
GradientBoostingRegressor	42%	1.3451	1.6904	34%	1.4535	1.7957
LGBMRegressor	72%	0.9204	1.1677	42%	1.3490	1.6880
MLPRegressor	34%	1.4399	1.8035	32%	1.4848	1.8256
SVR	31%	1.4414	1.8389	31%	1.4803	1.8446
keras	3%	1.7167	2.1885	1%	1.7359	2.2070

Table 18: The results of the machine learning models of data set number 2 (without the lithology readings).

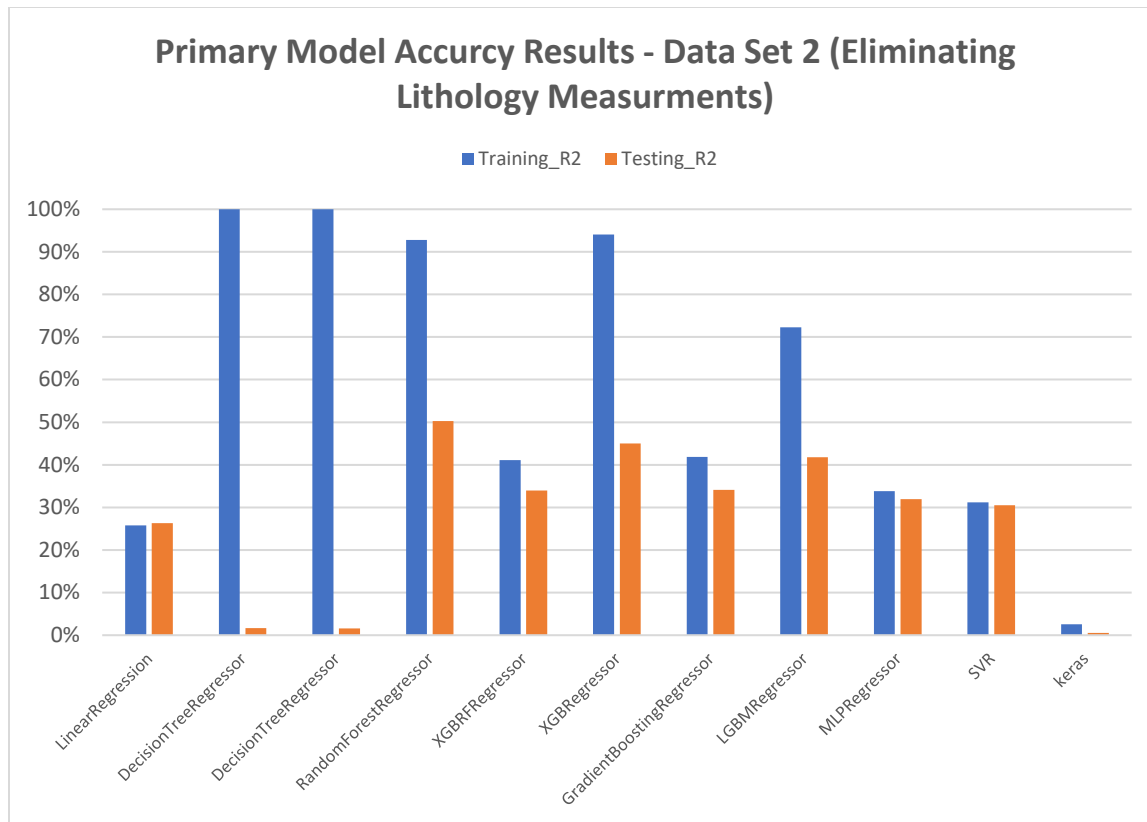


Figure 34: Accuracy comparison of training and testing results for the primary machine learning models of data set number 2 (without the lithology readings).

Results Remarks:

- Generally, the results of the machine learning models for data set number 2 after removing the lithology measurements are similar to the results before removing the lithology measurements. This supports the elimination of the lithology measurements in prediction of the total NMR porosity.
- After eliminating the lithology measurements, the three models Random Forest Regressor, XGBRegressor, and LGBMRegressor are still providing better results of the correlation coefficient (R^2) for the training and the testing groups comparing to the other machine learning models.

- Generally, the results of the correlation coefficient (R^2), the absolute average percentage error (AAPE), and root mean square error (RMSE) are still not sufficient. The required action will be increasing the machine learning models capabilities to improve the prediction results of NMR total porosity.

- Compiled Data Set:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	27%	1.5274	1.9111	25%	1.5325	1.9116
DecisionTreeRegressor	100%	0.0000	0.0000	67%	0.6644	1.2775
DecisionTreeRegressor	100%	2.93E-16	1.17E-15	65%	0.6672	1.3093
RandomForestRegressor	97%	0.2509	0.3628	80%	0.6852	0.9822
XGBRFRegressor	38%	1.4019	1.7584	35%	1.4291	1.7855
XGBRegressor	80%	0.7545	0.9915	63%	1.0354	1.3399
GradientBoostingRegressor	41%	1.3794	1.7274	37%	1.4098	1.7598
LGBMRegressor	59%	1.1452	1.4377	49%	1.2563	1.5801
MLPRegressor	33%	1.4631	1.8286	31%	1.4789	1.8395
SVR	33%	1.4387	1.8348	31%	1.4617	1.8427
keras	27%	1.5241	1.9084	25%	1.5369	1.9140

Table 19: The results of the machine learning models of the compiled data set.

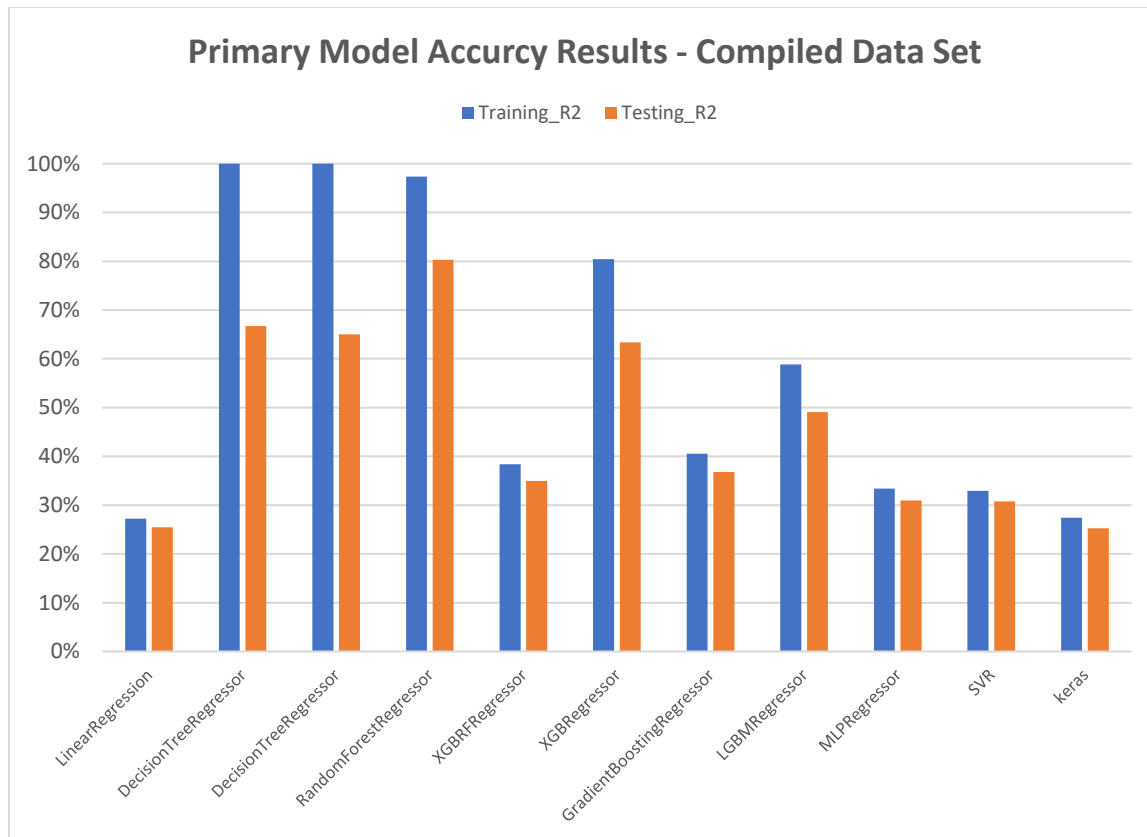


Figure 35: Accuracy comparison of training and testing results for the primary machine learning models of the compiled data set.

Results Remarks:

- Overall, the results from the machine learning models have provided higher values of the correlation coefficient (R^2) with the compiled data set in comparisons with the other data sets. This indicates that the increment in the number of data points can enhance the performance of the machine learning models.
- Several models have provided good results of the correlation coefficient (R^2) for the testing group such as Random Forest Regressor (80%), Decision Tree Regressor (67% & 65%), and XGBRegressor (63%). These results from the primary models

shows indications of the capability of machine learning models to predict total NMR porosity.

- The results of the models with the complied data set display cases of underfitting and overfitting once compared to figure-29. For this reason, more regularizations and complexity are required to enhance the results of complied data models.

5.3.2 Optimized Models (Model-B)

As the results of the primary models were not sufficient, an enhancement for the machine learning models is required to improve the model's accuracy. In order to enhance the model's prediction values, the machine learning models were optimized by applying hyperparameters tuning for each data set using the same machine learning algorithm that were used for the primary models.

- Data Set Number 1:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	28%	1.5175	1.9002	27%	1.5170	1.9002
DecisionTreeRegressor	26%	1.5454	1.9277	25%	1.5404	1.9300
DecisionTreeRegressor	21%	1.5911	1.9838	19%	1.6114	2.0045
RandomForestRegressor	100%	8.25E-15	1.21E-14	69%	0.9381	1.2437
XGBRFRegressor	8%	1.7233	2.1396	7%	1.7326	2.1525
XGBRegressor	85%	0.6815	0.8804	57%	1.1528	1.4702
GradientBoostingRegressor	41%	1.3669	1.7139	38%	1.4006	1.7610
LGBMRegressor	62%	1.0832	1.3741	48%	1.2665	1.6028
MLPRegressor	95%	0.3650	0.4973	32%	1.3737	1.8332
SVR	55%	1.0481	1.5017	42%	1.2725	1.6938
keras	28%	1.5158	1.9023	27%	1.5183	1.9063

Table 20: The results of the machine learning models of data set number 1 from the optimized model.

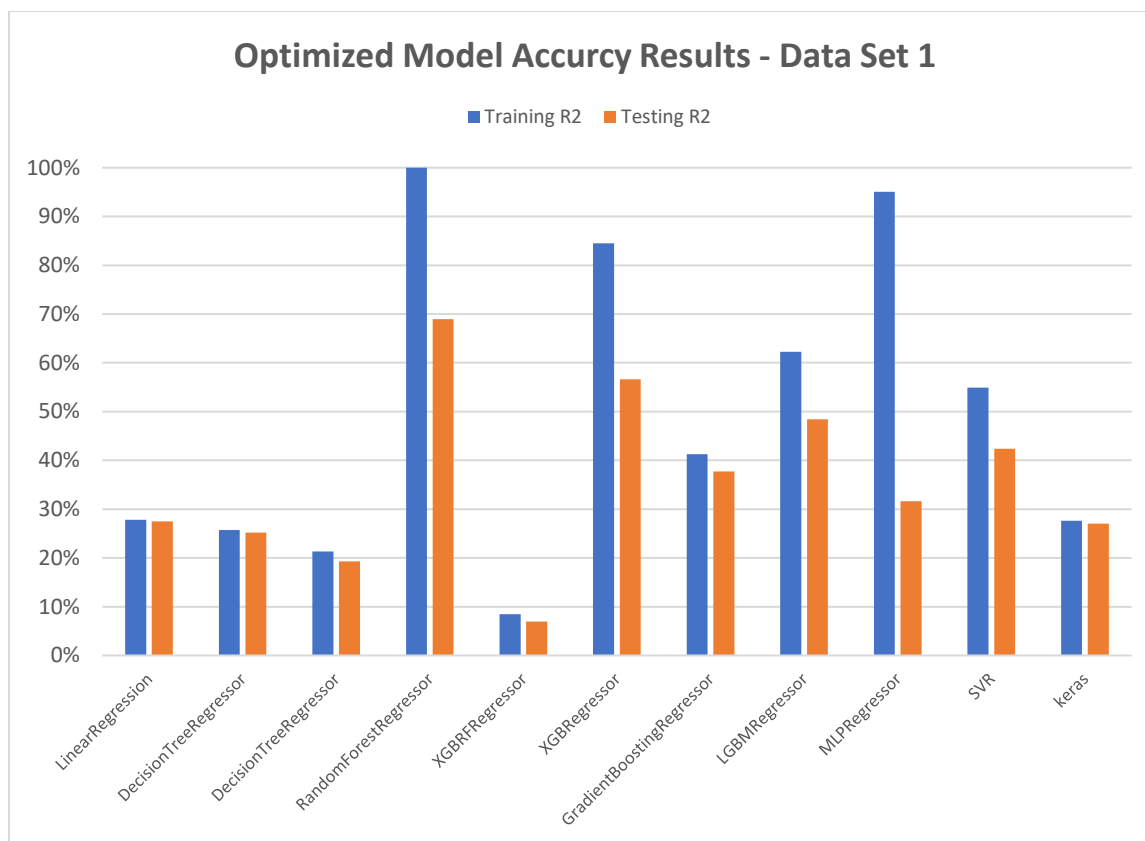


Figure 36: Accuracy comparison of training and testing results for the optimized machine learning models of data set number 1.

Results Remarks:

- From table-20 and figure-34, it can be obtained that the process of the machine learning models optimization (hyperparameters selection) has led to changes in the results of the correlation coefficient (R^2), absolute average percentage error (AAPE), and root mean square error (RMSE). Some of the algorithms have provided better results for both the training and the testing groups and others have encountered a decline in the models' accuracy.
- The Random Forest Regressor, XGBRegressor, and LGBMRegressor algorithms have exhibited minor enhancements of the total NMR prediction accuracy. These algorithms delivered higher values of the correlation coefficient (R^2) for both

training and testing groups (100% training 69% testing for Random Forest Regressor, 85% training 57% testing for XGBRegressor, 62% training 48% testing for LGBMRegressor). Moreover, the produced error values (AAPE and RMSE) from the models were reduced.

- MLPRegressor has delivered higher numbers for the correlation coefficient (R^2) of the training group and provided lower values for the testing groups. This indicates to overfitting of the model.
- The XGBRFRegressor algorithm has provided the lowest values of the correlation coefficient and the highest error values among the other algorithms.
- Overall, the machine learning models have lower accuracy for both training and testing groups. This issue might be addressed by either adding more data points (the compiled data set) or adding more features (using clustering algorithms).

- Data Set Number 2:

From the results of the primary model, it was interpreted that the results with the lithology variables and after eliminating the lithology variables are similar. For that, the further results of the machine learning model will only consider the variables of data set number 2 after eliminating the lithology variables. This step should accelerate the running time of the models.

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	26%	1.5288	1.9102	26%	1.5350	1.8997
DecisionTreeRegressor	24%	1.5429	1.9287	24%	1.5608	1.9307
DecisionTreeRegressor	25%	1.54E+00	1.92E+00	24%	1.5434	1.9276
RandomForestRegressor	100%	8.23E-15	0.0000	54%	1.1990	1.5071
XGBRFRegressor	10%	1.6868	2.0990	7%	1.7171	2.1320
XGBRegressor	90%	0.5415	0.7003	48%	1.2768	1.5948
GradientBoostingRegressor	100%	0.0065	0.0088	56%	1.1603	1.4742
LGBMRegressor	77%	0.8305	1.0688	43%	1.3346	1.6748
MLPRegressor	99%	0.1708	0.2310	24%	1.4726	1.9334
SVR	90%	0.3403	0.7041	51%	1.1844	1.5552
keras	18%	1.5980	2.0051	18%	1.6086	2.0096

Table 21: The results of the machine learning models of data set number 2 (without the lithology variables).

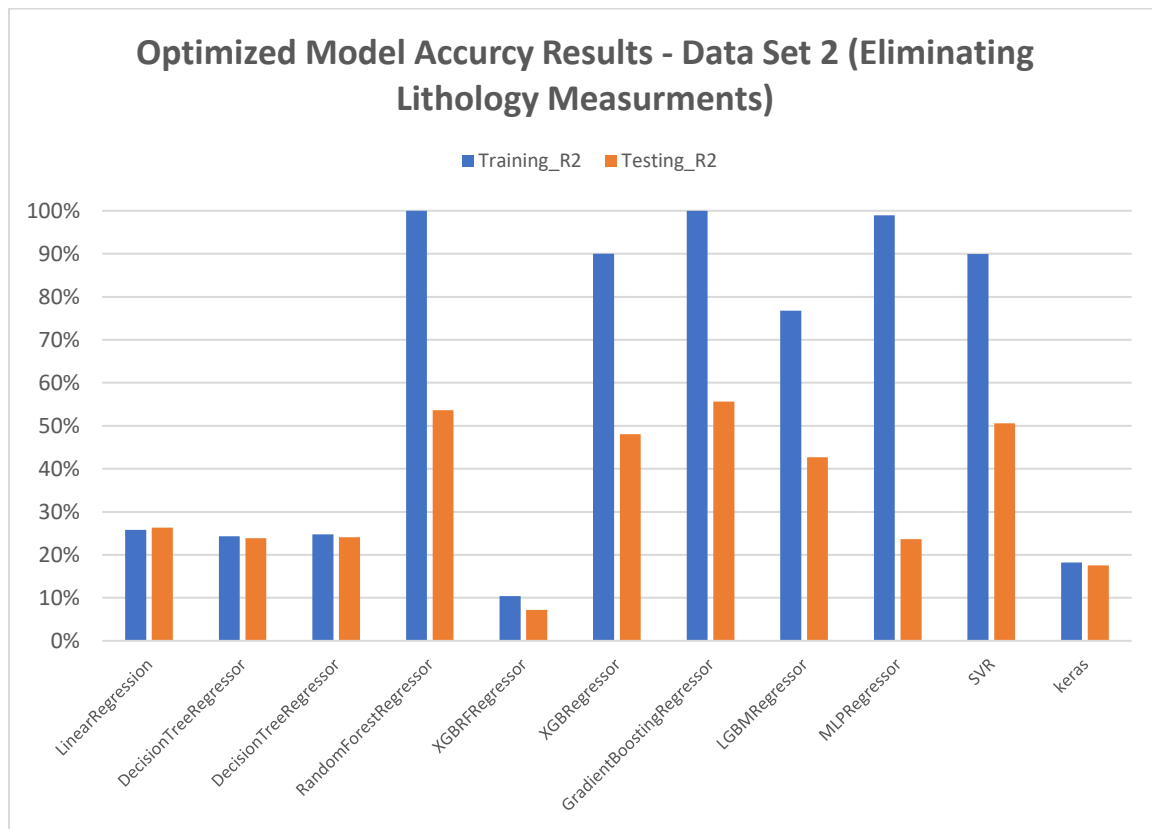


Figure 37: Accuracy comparison of training and testing results for the optimized machine learning models of data set number 2 (without the lithology variables).

Results Remarks:

- Generally, the results of predicting total NMR porosity were improved for all algorithms in the optimized machine learning model (Model-B) of data set number 2 for both training and testing groups.
- The optimized model of data set number 2 has provided similar results to the optimized model of data set number 1, where Random Forest Regressor, XGBRegressor, and LGBMRegressor algorithms besides GradientBosstingRegressor algorithm have delivered the highest values of correlation coefficient. In addition to these algorithms, Support Vector Machine Regressor (SVR) has provided higher values for the correlation coefficient for both training and testing groups in comparison with the primary model.
- Similar to data set number 1, the XGBRFRegressor has delivered the lowest values of the correlation coefficient and the highest numbers for both AAPE and RMSE. This includes the training and testing group, where the delivered correlation coefficient values are 10% and 7%, respectively.
- In total, the results for data set number 1 and data set number 2 seem to be similar for the most regression algorithms. These outcomes direct the study towards the same analysis direction of adding more data points and testing the clustering algorithms to enhance the machine learning models' results.

- Compiled Data Set:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	27%	1.5274	1.9111	25%	1.5325	1.9116
DecisionTreeRegressor	20%	1.6088	2.0045	20%	1.5955	1.9839
DecisionTreeRegressor	24%	1.57E+00	1.96E+00	21%	1.5689	1.9667
RandomForestRegressor	100%	0.0000	0.0000	82%	0.5451	0.9475
XGBRFRegressor	8%	1.7279	2.1435	7%	1.7130	2.1331
XGBRegressor	90%	0.5430	0.7158	69%	0.9329	1.2238
GradientBoostingRegressor	100%	0.0992	0.1315	87%	0.5034	0.7921
LGBMRegressor	62%	1.0874	1.3740	52%	1.2185	1.5362
MLPRegressor	94%	0.3752	0.5268	56%	0.9773	1.4699
SVR	56%	1.0291	1.4912	43%	1.2191	1.6682
keras	27%	1.5338	1.9192	25%	1.5444	1.9224

Table 22:: The results of the machine learning models of the compiled data set.

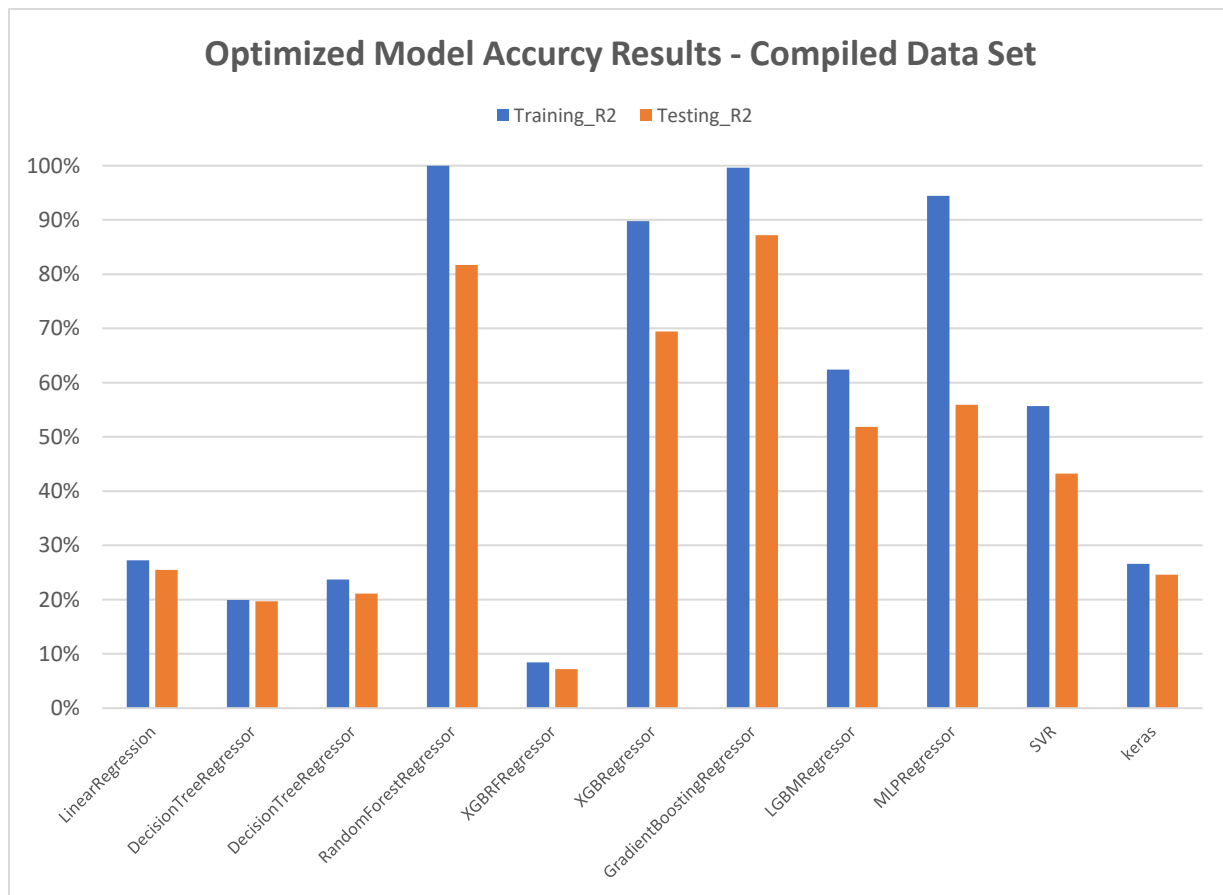


Figure 38: Accuracy comparison of training and testing results for the optimized machine learning models of the compiled data set.

Results Remarks:

- The optimization process for the primary machine learning model (Model-A) raised outcoming values of the correlation coefficient and reduced the error values for many of the used algorithms with the complied data set.
- The GradientBoostingRegressor algorithms has exhibited a significant increment in the optimized model (Model-B) accuracy, where it provided the values of 100% for the training and 87% for the testing of the correlation coefficient. Moreover, the error values have been reduced as all the numbers of errors are less 1.00.
- Minor increments in the correlation coefficient values for Random Forest Regressor and XGBRegressor algorithms has been resulted from Model-B, where the training group values are 100% and 90% and the testing group values are 82% and 69%, respectively. In addition, both error matrices AAPE and RMSE values have been slightly reduced in comparison to Model-A.
- Generally, the additional number of data points has improved the machine learning model accuracy. Further enhancement to the machine learning model including the increasing the model complexity and adding more features by using clustering algorithms can be applied to increase the model's accuracy.

5.3.2 Models with Clustered Data (Model-C)

As the results of increasing the model complexity were not sufficient to deliver the required results, further developments of the model and the data analysis are needed. For this reason, data clustering was used to add more features, which are the number of clusters. This means

that the classification of the data points will be utilized to expedite the training process, where the data points with similar features will be allocated within the same cluster.

Several clustering algorithms were used to distribute the data into different clusters, which provided different number of clusters for each algorithm. The clustering process of these algorithms were evaluated by using the Silhouette Coefficient (Silhouette Score), where the average score for all data points was computed as part of the python code. In addition, to ensure the selection of the best parameters for each clustering algorithm, hyperparameters tuning was conducted on each algorithm as part of the data clustering process.

- Data Set Number 1:

The below table-23 shows the best average Silhouette score and the number of clusters for each clustering algorithm using the data set number-1.

Clustering Algorithm	Best Average Silhouette Score	Number of Clusters
K-means	0.45	2
DBSCAN	0.31	5
Affinity Propagation	0.18	251
BIRCH	0.50	3
Mean Shift	0.64	2

Table 23: The results of the clustering algorithms using data set number 1.

From table-23, it can be seen that the Mean shift algorithm is displaying the best Silhouette score with separating the data into two (2) clusters. This indicates that the best clustering algorithm for data set number 1 is Mean Shift algorithm.

By using the provided number of clusters in table-23 from the Mean Shift algorithm, the machine learning models were trained and tested. The table-24 below shows the results of the machine learning model with the clustered data.

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	28%	1.5146	1.8976	28%	1.5138	1.8958
DecisionTreeRegressor	25%	1.5440	1.9323	23%	1.5640	1.9522
DecisionTreeRegressor	25%	1.5509	1.9406	24%	1.5542	1.9418
RandomForestRegressor	100%	8.25E-15	1.21E-14	70%	0.9247	1.2296
XGBRFRegressor	8%	1.7247	2.1412	7%	1.7342	2.1548
XGBRegressor	84%	0.6941	0.8951	55%	1.1705	1.4913
GradientBoostingRegressor	100%	0.0951	0.1233	77%	0.8084	1.0585
LGBMRegressor	62%	1.0858	1.3758	47%	1.2801	1.6196
MLPRegressor	96%	0.3428	0.4723	37%	1.3307	1.7702
SVR	55%	1.0481	1.5017	42%	1.2729	1.6945
keras	28%	1.5130	1.8993	27%	1.5151	1.9025

Table 24: The results of the machine learning models of data set number 1 after adding the clustering features.

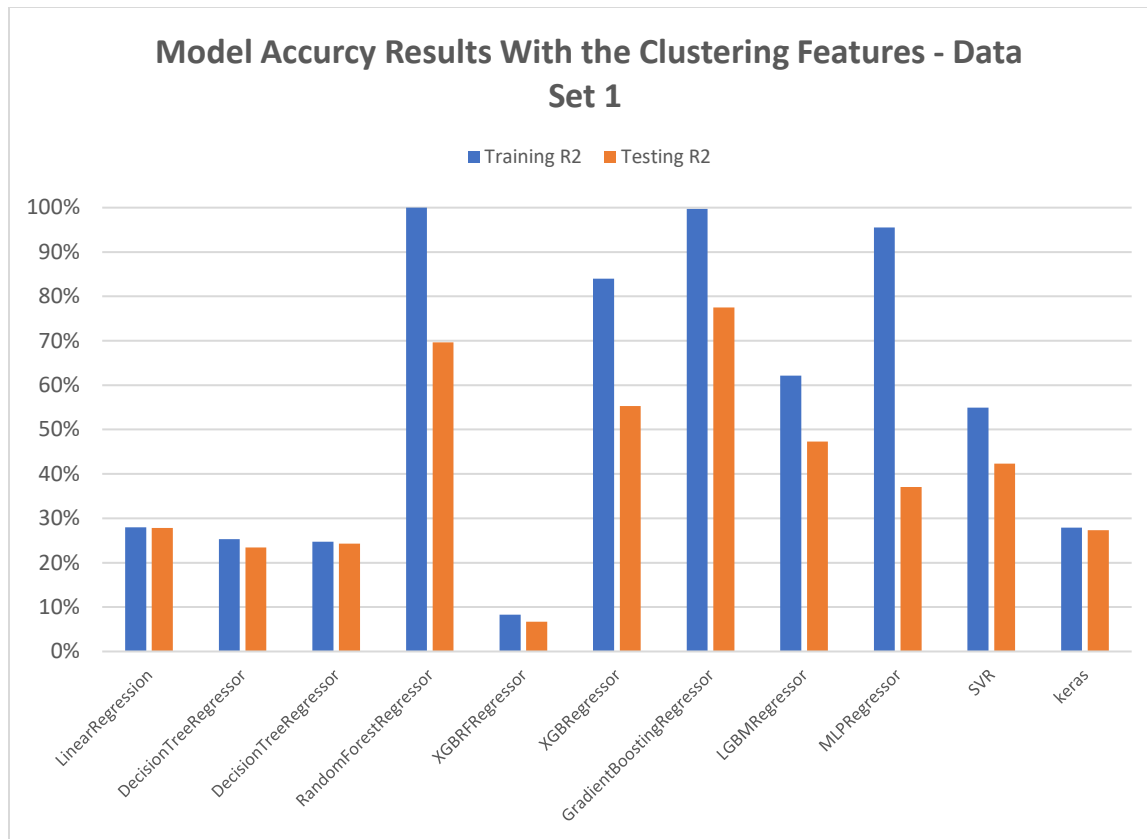


Figure 39: Accuracy comparison of training and testing results for the machine learning models that include clustering features of data set 1.

Results Remarks:

- By including the clustering features as input variables in Model-C for data set number 1, the Gradient Boosting Regressor algorithm has delivered the highest values of correlation coefficient and the lowest error values among the other algorithms. The algorithm has provided the results of 100% for training and 77% for the testing groups for the correlation coefficient.
- The Random Forest Regressor algorithm has added minor improvement to the accuracy of the model for the testing group, where is the model accuracy reached 70%. In addition, the error values were slightly reduced to 0.9247 for the AAPE and 1.2296 for the RMSE. As for the training group, the results from Model-B are

still maintained. This required for more adjustment to the model to deliver closed results for both groups.

- Overall, the results of the other machine learning algorithms did not show any major improvement or reduction on the prediction accuracy from Model-B. This will require to review the model hyperparameters and configuration to approach better numbers.

- Data Set Number 2:

By running the same clustering algorithms on data set number 2, the values of average Silhouette score and the number of clusters have been produced as displayed in table-25.

Clustering Algorithm	Best Average Silhouette Score	Number of Clusters
K-means	0.45	2
DBSCAN	Only 1 cluster (Not valid)	
Affinity Propagation	0.51	11
BIRCH	0.57	2
Mean Shift	0.87	2

Table 25: The results of the clustering algorithms using data set number 2.

Similar to data set number 1, the Mean Shift algorithm have provided the highest values for Silhouette score. In addition, data set number 2 was classified into two (2) clusters similar to data set number 1. As the new features of data clustering have been added to data set number 2, the machine learning models have been tested and produced the below values for training and testing groups of data (table-26).

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	26%	1.5286	1.9102	26%	1.5347	1.8996
DecisionTreeRegressor	21%	1.5811	1.9728	20%	1.6014	1.9808
DecisionTreeRegressor	20%	1.5847	1.9809	20%	1.6054	1.9809
RandomForestRegressor	100%	8.23E-15	1.29E-14	54%	1.1905	1.5036
XGBRFRegressor	10%	1.6867	2.0987	7%	1.7168	2.1319
XGBRegressor	96%	0.3219	0.4246	49%	1.2480	1.5757
GradientBoostingRegressor	42%	1.3451	1.6904	34%	1.4533	1.7954
LGBMRegressor	77%	0.8249	1.0632	43%	1.3353	1.6665
MLPRegressor	99%	0.1462	0.1943	28%	1.4367	1.8815
SVR	90%	0.3403	0.7041	51%	1.1844	1.5552
keras	6%	1.6844	2.1440	5%	1.6815	2.1538

Table 26: The results of the machine learning models of data set number 2 after adding the clustering features.

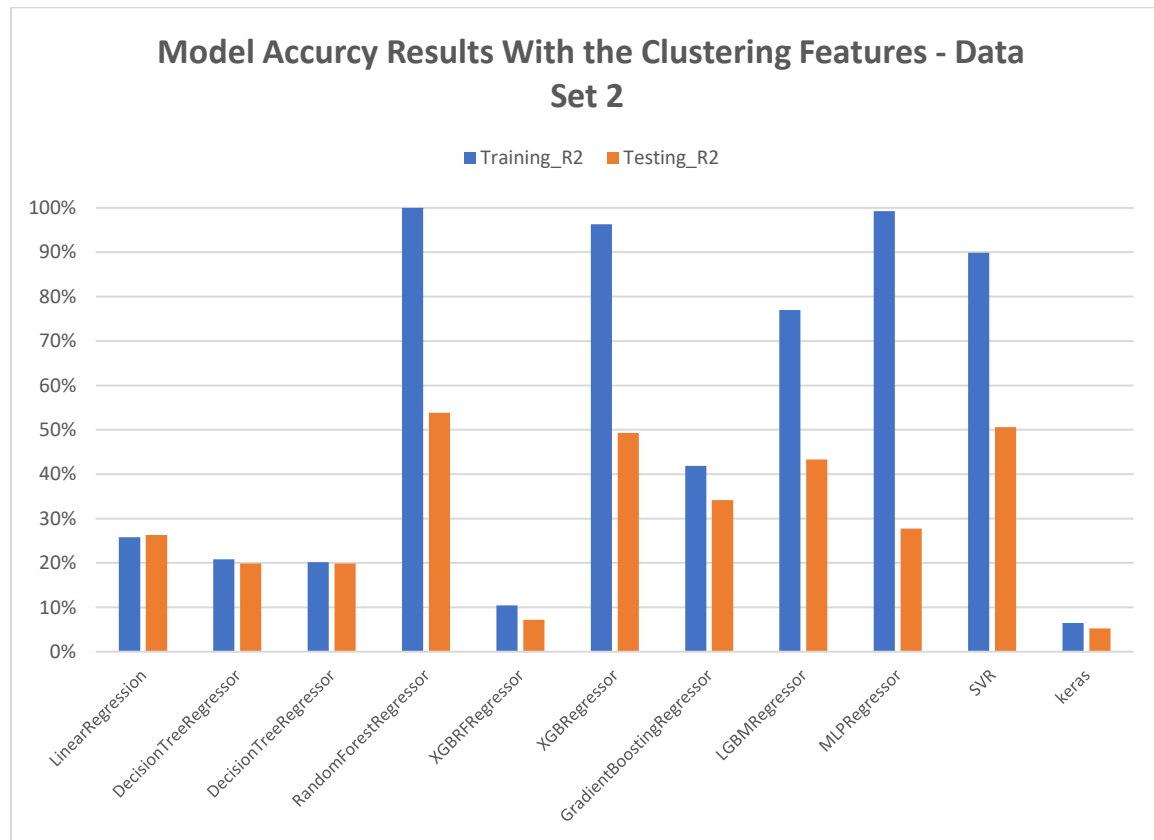


Figure 40: Accuracy comparison of training and testing results for the machine learning models that include clustering features of data set 1.

Results Remarks:

- Generally, adding the clustering features to data set number 2 as input variables did not illustrate any major improvements from the results of Model-B for any of the machine learning algorithms.
- There are two exceptions where XGBRegressor has enhanced the training process to achieve 96% accuracy for the training group. The other model was MLP Regressor, where the testing accuracy approached 28%. However, the results of these algorithms are still demonstrating large difference between the training and the testing results, which is an indication of model overfitting.
- The Gradient Boosting Regressor and Keras regressor has displayed a major decline of the accuracy from Model-B for both the training and the testing groups. This requires an extensive change in the model complexity to enhance the results.
- All the results of data set number 2 from the algorithms in Model-C are not satisfactory and further action of model complexity enhancements are required to deliver the desired results.

- Complied Data Set:

Table-27 is displaying the average Silhouette score and the number of clusters for the complied data set using the different clustering algorithms.

Clustering Algorithm	Best Average Silhouette Score	Number of Clusters
K-means	0.45	2
DBSCAN	0.13	2
Affinity Propagation	Was not computed due to high RAM requirements	
BIRCH	0.51	3
Mean Shift	0.64	2

Table 27: The results of the clustering algorithms using data set number 1.

The Mean Shift algorithm has provided the best values of Silhouette score, which is a similar case for data set number 1 and data set number 2. The resulted number of clusters for the complied data set is two (2) clusters. Using the data labels of clustering, the results of running the machine learning models using have produced and inserted into table-28.

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	27%	1.5251	1.9087	26%	1.5287	1.9078
DecisionTreeRegressor	19%	1.6113	2.0108	19%	1.6070	1.9958
DecisionTreeRegressor	26%	1.5382	1.9233	25%	1.5310	1.9218
RandomForestRegressor	100%	8.39E-15	1.26E-14	80%	0.5756	0.9882
XGBRFRegressor	9%	1.7248	2.1406	7%	1.7105	2.1307
XGBRegressor	85%	0.6783	0.8773	66%	1.0056	1.2982
GradientBoostingRegressor	100%	0.0998	0.1320	87%	0.5059	0.7964
LGBMRegressor	62%	1.0962	1.3857	51%	1.2228	1.5427
MLPRegressor	94%	0.3956	0.5622	58%	0.9681	1.4334
SVR	56%	1.0291	1.4912	43%	1.2191	1.6682
keras	26%	1.5343	1.9279	24%	1.5414	1.9288

Table 28: The results of the machine learning models of the complied data set after adding the clustering features.

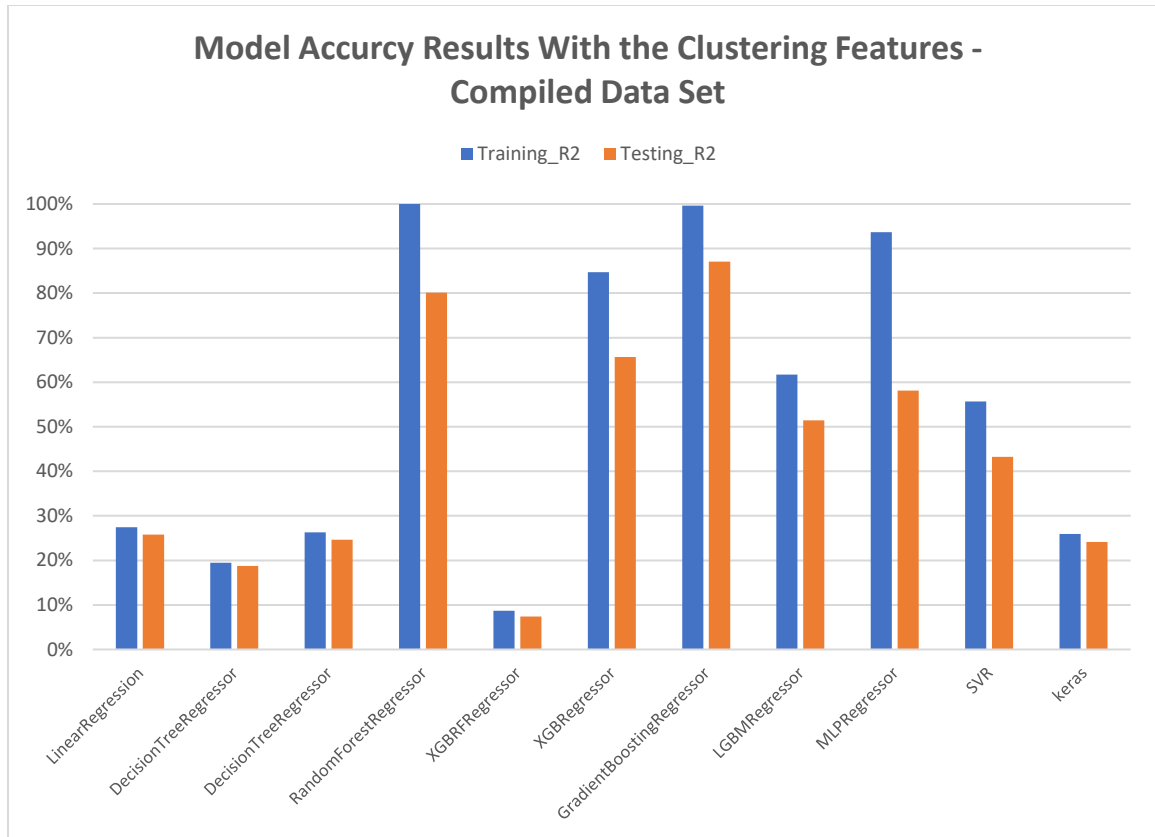


Figure 41: Accuracy comparison of training and testing results for the machine learning models that include clustering features of the compiled data set.

Results Remarks:

- Comparing the results of Model-B and Model-C displays that the additional clustering features did not participate in improving the model prediction accuracy as there were only minor (positive or negative) changes in Model-C results.
- An investigation of the hyperparameters of the used algorithms is required to confirm the model results.

5.3.4 Optimized Models with Clustered Data (Model-D)

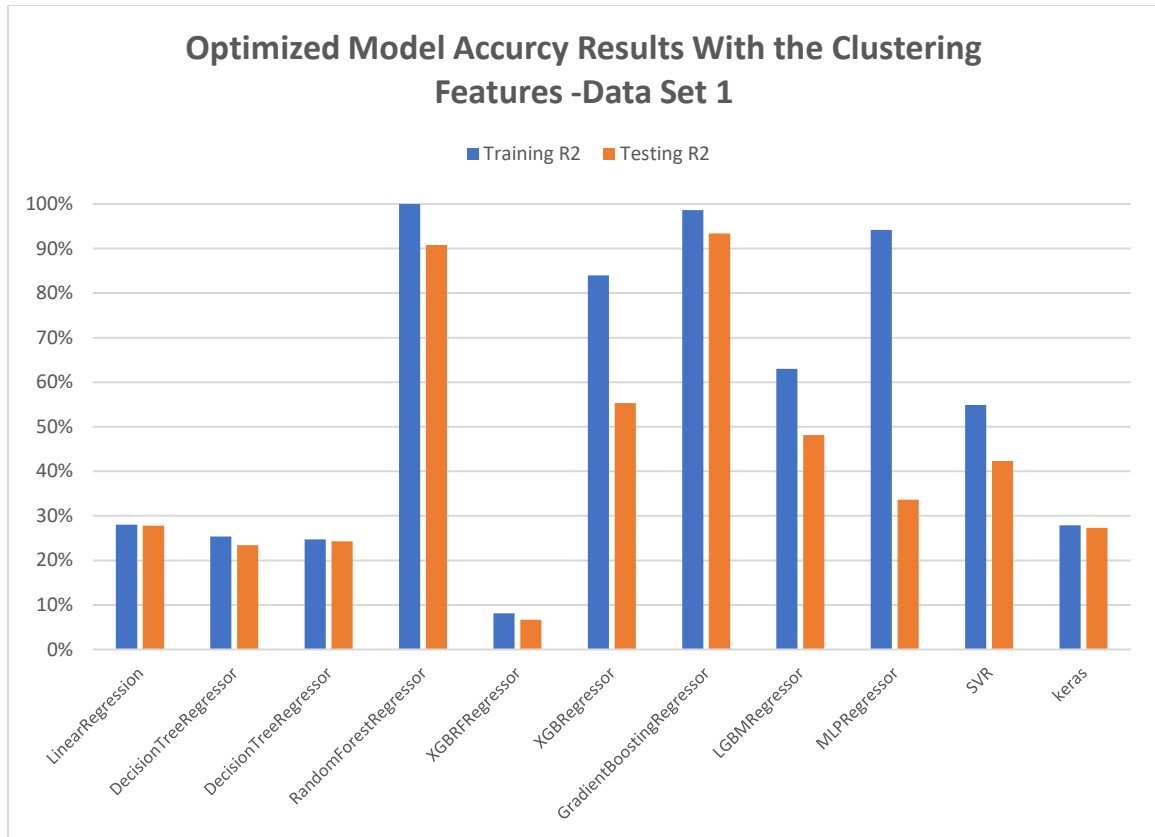
The results of Model-C, which included new clustering features, did not add any major improvements to the accuracy of the total porosity prediction model. This could be related

to the model configuration, where it might not be properly fit with the data after adding the new features. One approach to overcome this issue to re-tune the model by selecting the best hyperparameters. The process of hyperparameter tuning is applied on each data set separately to find the most suitable combination of hyperparameters.

- Data Set Number 1:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	28%	1.5146	1.5146	28%	1.5138	1.8958
DecisionTreeRegressor	25%	1.5440	1.9323	23%	1.5640	1.9522
DecisionTreeRegressor	25%	1.5509	1.9406	24%	1.5542	1.9418
RandomForestRegressor	100%	1.01E-05	0.0012	91%	0.2842	0.6782
XGBRFRegressor	8%	1.7256	2.1434	7%	1.7343	2.1551
XGBRegressor	84%	0.6941	0.8951	55%	1.1705	1.4913
GradientBoostingRegressor	99%	0.1966	0.2587	93%	0.3960	0.5771
LGBMRegressor	63%	1.0776	1.3605	48%	1.2730	1.6072
MLPRegressor	94%	0.3894	0.5392	34%	1.3689	1.8177
SVR	55%	1.0481	1.5017	42%	1.2729	1.6945
keras	28%	1.5130	1.8993	27%	1.5151	1.9025

Table 29: The results of the machine learning models of data set number 1 (including the clustering features) in a new optimized model.



Results Remarks:

- By running the tuning process for the machine learning algorithm to adjust the results from Model-C, most algorithms provided similar results in Model-D as in Model-C. The only two expectations were Random Forest Regressor algorithm and Gradient Boosting Regressor Algorithm.
- Random Forest Regressor algorithm has achieved a training accuracy of 100% for the correlation coefficient and minor error values of $1.01e-5$ for the AAPE and 0.0012 for the RMSE. As for testing the model, it managed to reach 91% score of the correlation coefficient, 0.2842 for the AAPE, and 0.6782 for the RMSE.
- The Gradient Boosting Regressor algorithm has delivered 99% of correlation coefficient, 0.1966 for the AAPE, and 0.2587 for the RMSE for the training group.

In addition, it provided 93% accuracy on the correlation coefficient measurement with 0.3960 for the AAPE and 0.5771 for the RMSE for the testing group.

- Data Set Number 2:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	26%	1.5219	1.8968	25%	1.5452	1.9275
DecisionTreeRegressor	22%	1.5575	1.9467	20%	1.5974	1.9879
DecisionTreeRegressor	22%	1.5597	1.9459	21%	1.5985	1.9822
RandomForestRegressor	100%	8.23E-15	1.29E-14	54%	1.1930	1.5048
XGBRFRegressor	11%	1.6786	2.0868	9%	1.7147	2.1310
XGBRegressor	90%	0.5358	0.6938	49%	1.2668	1.5868
GradientBoostingRegressor	100%	0.0736	0.0953	56%	1.1535	1.4673
LGBMRegressor	77%	0.8305	1.0688	43%	1.3346	1.6748
MLPRegressor	98%	0.2436	0.3311	33%	1.3872	1.8069
SVR	90%	0.3403	0.7041	51%	1.1844	1.5552
keras	6%	1.6844	2.1440	5%	1.6815	2.1538

Table 30: The results of the machine learning models of data set number 2 (including the clustering features) in a new optimized model.

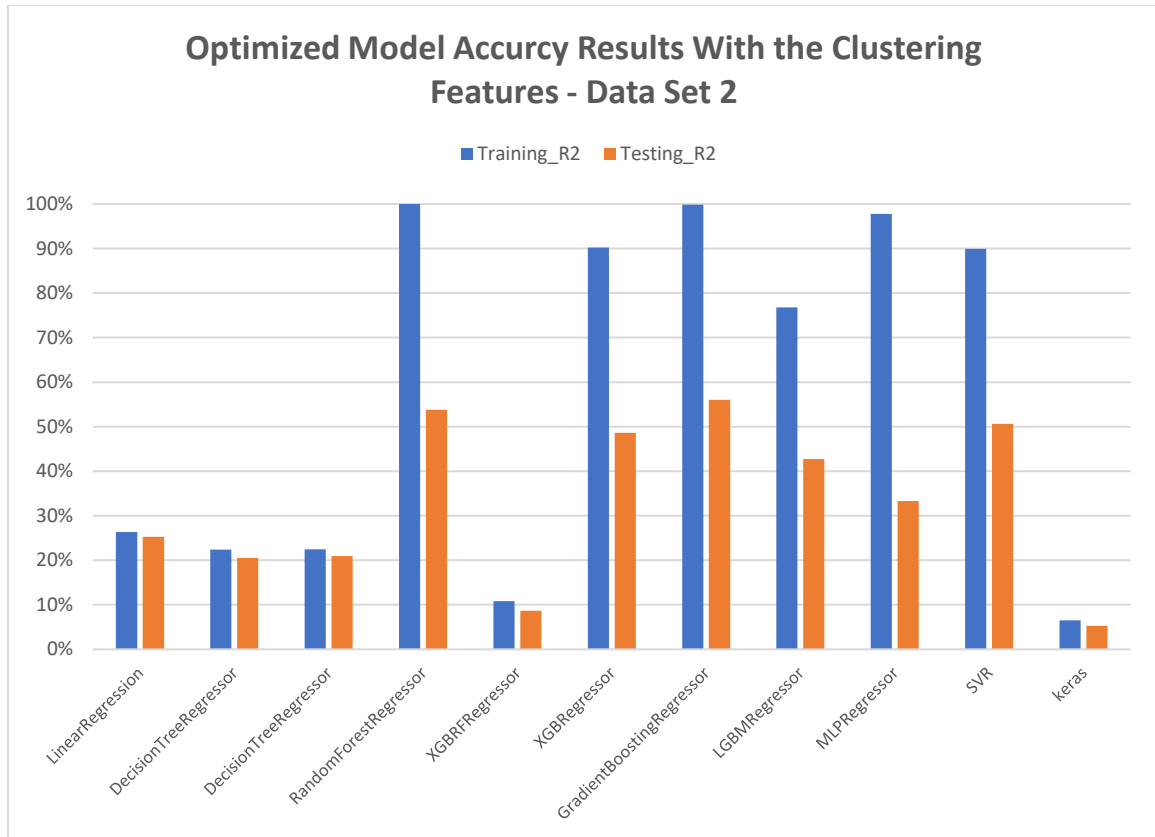


Figure 42: Accuracy comparison of training and testing results for the optimized machine learning models after including clustering features with set number 2.

Results Remarks:

- The tuning of the hyperparameters using data set number 2 has delivered similar results for all the machine learning algorithms comparing to Model-C. There only improvements in prediction accuracy for the results of the Gradient Boosting Regressor Algorithm and MLPRegressor.
- Data set number 2 did not deliver any acceptable results from the 4 models (A, B, C, and D). Further data analysis might be required before re-testing the data set with the machine learning models.

- Compiled Data Set:

Model	Training R2	Training AAPE	Training RMSE	Testing R2	Testing AAPE	Testing RMSE
LinearRegression	27%	1.5286	1.9106	27%	1.5186	1.9030
DecisionTreeRegressor	20%	1.5978	1.9931	20%	1.5942	1.9918
DecisionTreeRegressor	24%	1.5646	1.9474	24%	1.5539	1.9428
RandomForestRegressor	99%	0.0948	0.1716	95%	0.2607	0.4776
XGBRFRegressor	8%	1.7238	2.1392	8%	1.7171	2.1368
XGBRegressor	89%	0.5532	0.7275	69%	0.9394	1.2368
GradientBoostingRegressor	100%	0.1157	0.1568	97%	0.2177	0.3765
LGBMRegressor	61%	1.1102	1.3950	56%	1.1678	1.4713
MLPRegressor	93%	0.4180	0.5878	57%	0.9951	1.4520
SVR	58%	0.9817	1.4470	53%	1.0650	1.5262
keras	28%	1.5247	1.9010	28%	1.5137	1.8946

Table 31: The results of the machine learning models of the complied data set (including the clustering features) in a new optimized model.

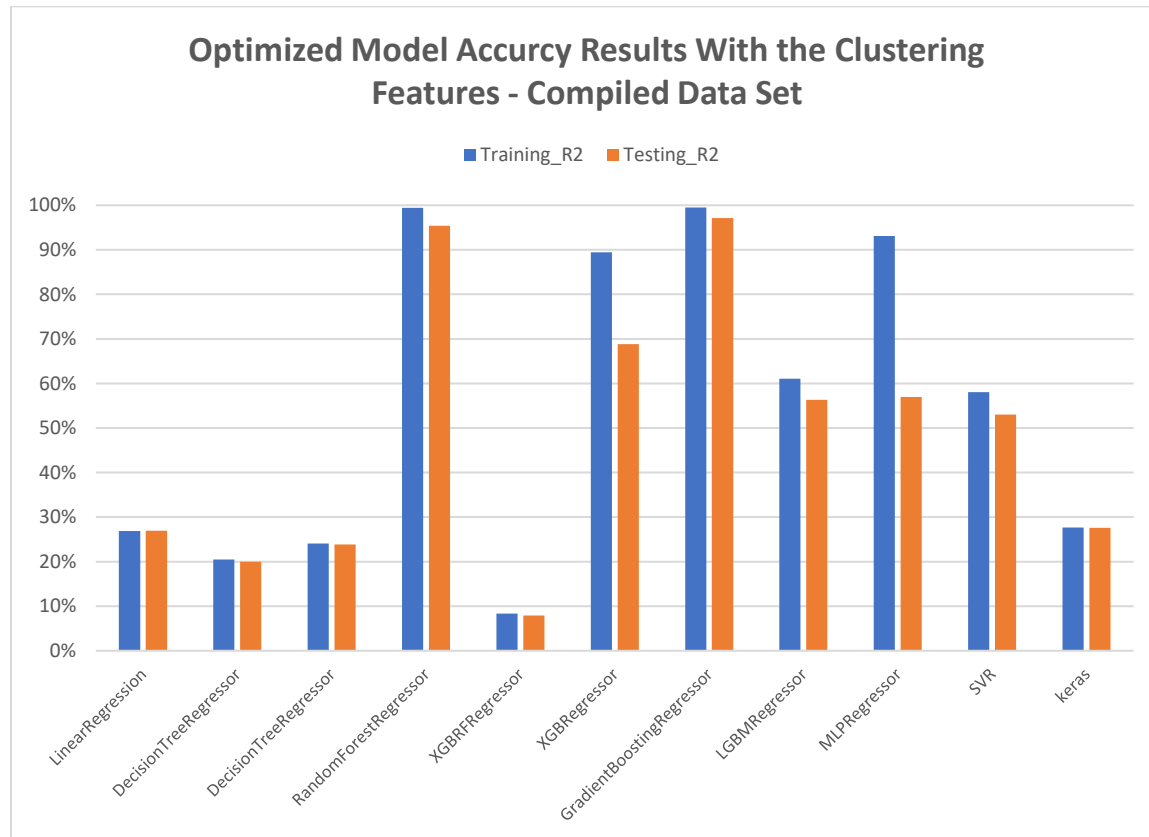


Figure 43: Accuracy comparison of training and testing results for the optimized machine learning models after including clustering features with the compiled set.

Results Remarks:

- The tuning of the hyperparameters of Model-C, which included the clustering features as additional input variables, results in increasing the prediction accuracy of Model-D for two algorithms. The two algorithms are Random Forest Regressor and Gradient Boosting Regressor. As for the other algorithms, there were displaying minor positive or negative changes in the prediction accuracy of Model-D.
- The Random Forest Regressor algorithm has provided the prediction results of 99% and 95% for the correlation coefficient, 0.0948 and 0.2607 for the AAPE, 0.1716 and 0.4776 for the RMSE for both of the training and the testing groups, respectively.
- The Gradient Boosting Regressor algorithm has exceeded the results of the Random Forest Regressor, where it provided the values of 100% and 97% for the correlation coefficient, 0.1157 and 0.2177 for the AAPE, 0.1568 and 0.3765 for the RMSE for both of the training and the testing groups, respectively.

Results Validation:

As part of the results check, a computed cross validation was conducted during the training process to eliminate the overfitting situation. The cross validation was computed using the below code, where five (5) splits were selected randomly and tested by the code. The validation scores are reported on the cases where the results of the algorithms were exceeding 90% of the correlation coefficient score for the testing group. As for the results of the cross validation of other algorithms, there were utilized during the model training to

tune the hyperparameters of the different algorithms. The results of the cross validation are reported in table.

```
scores_train_model_ = cross_val_score(model, X_train_scaled,
y_train, cv=5)
print("%0.2f accuracy with a standard deviation of %0.2f" %
(scores_train_model_rfr.mean()))
```

Model	Data Set	Algorithm	Training R2	Testing R2	Cross Validation R2
D	Data Set 1	RandomForestRegressor	100%	91%	86%
		GradientBoostingRegressor	99%	93%	90%
D	Compiled Data Set	RandomForestRegressor	99%	95%	92%
		GradientBoostingRegressor	100%	97%	96%

Table 32: The cross-validation results for Random Forest Regressor and Gradient Boosting Regressor algorithms.

The Figures from figure-42 to figure-49 display the results of the total NMR prediction values in comparison with the actual NMR values for the Random Forest Regressor algorithm and Gradient Boosting Regressor algorithm.

- Data Set 1:

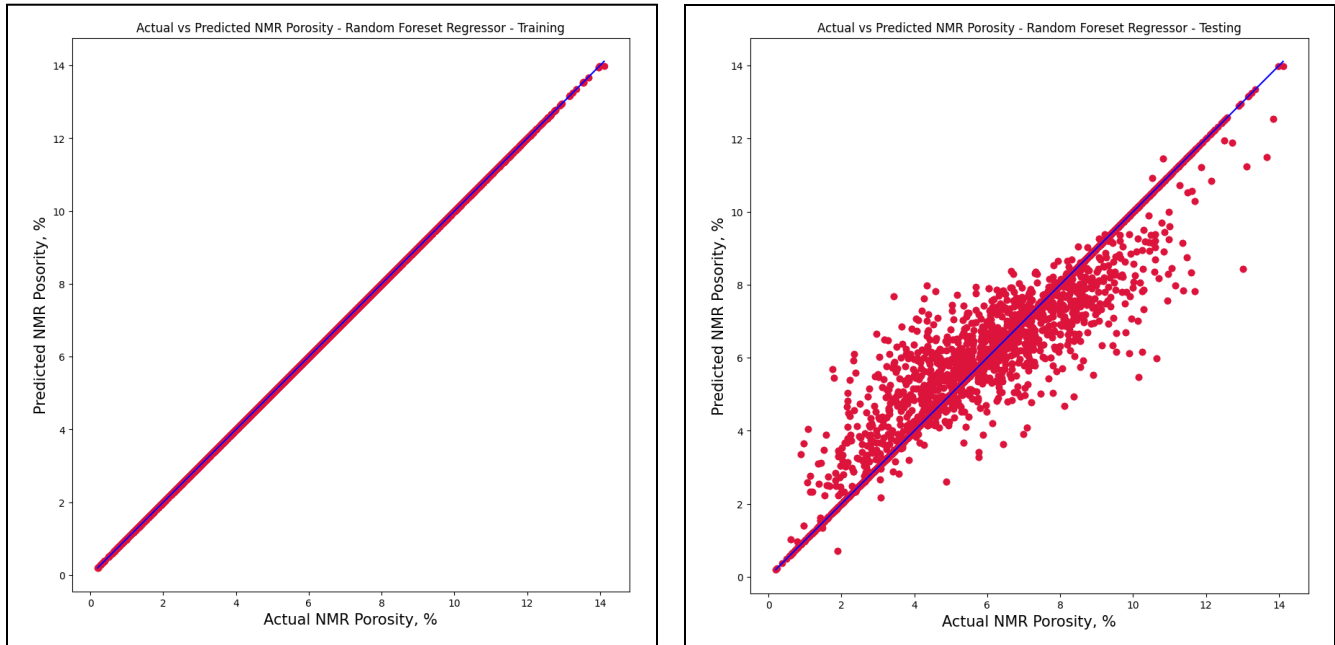


Figure 44: The correlation between the predicted values and the actual values of the total NMR porosity using the Random Forest Regressor algorithm for the training and testing groups of data set 1.

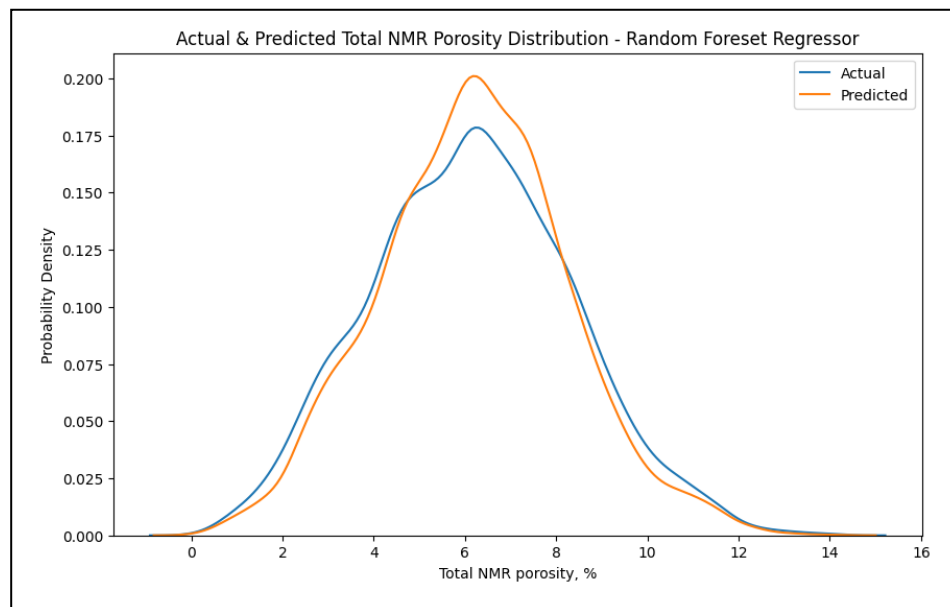


Figure 45: Distribution of actual and predicted values of total NMR porosity using Random Forest Regressor algorithm for testing group of set 1.

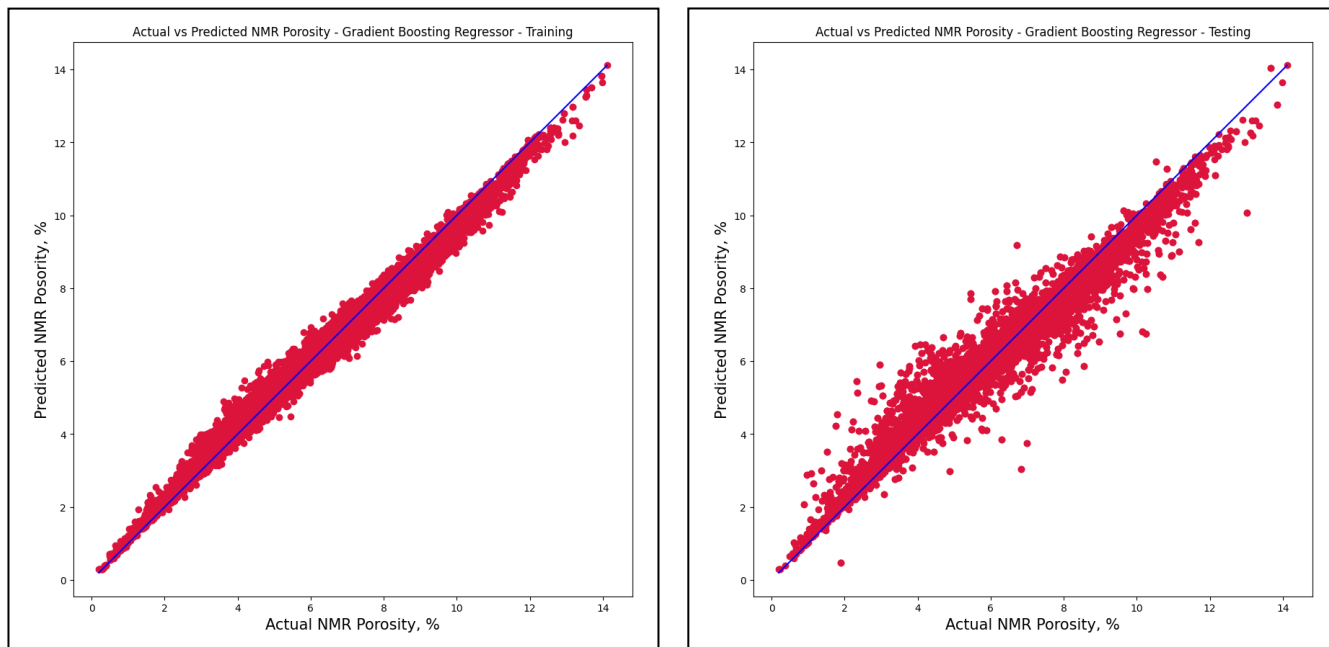


Figure 46: The correlation between the predicted values and the actual values of the total NMR porosity using the Gradient Boosting Regressor algorithm for the training and testing group of data set 1.

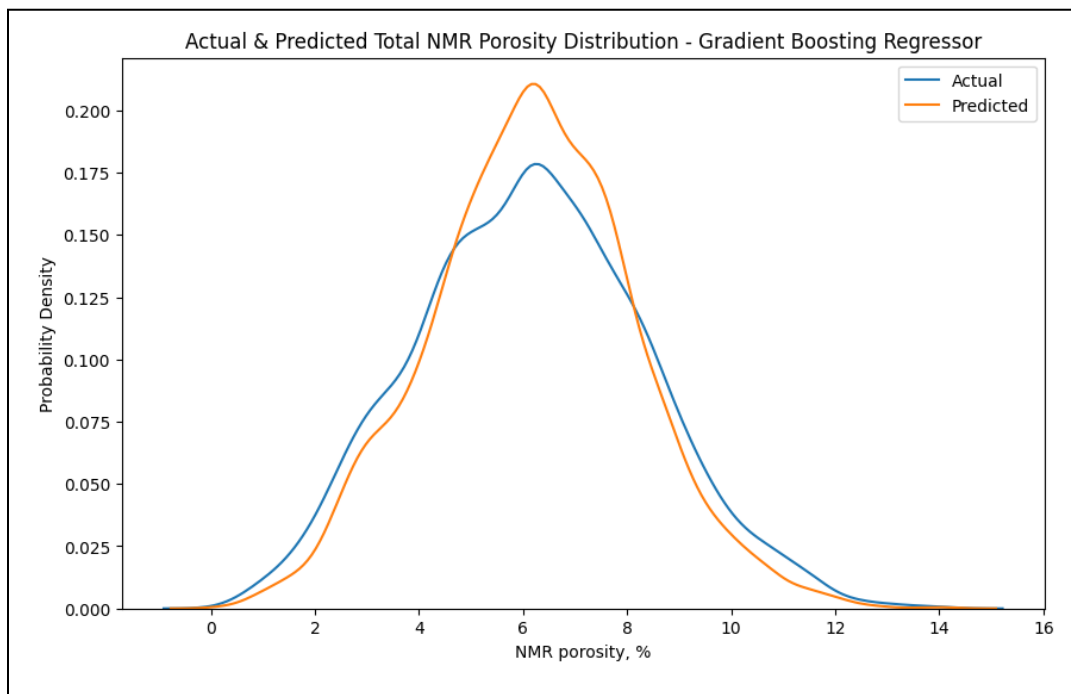


Figure 47: Distribution of actual and predicted values of total NMR porosity using Gradient Boosting Regressor algorithm for testing group of data set 1.

- Compiled Data Set:

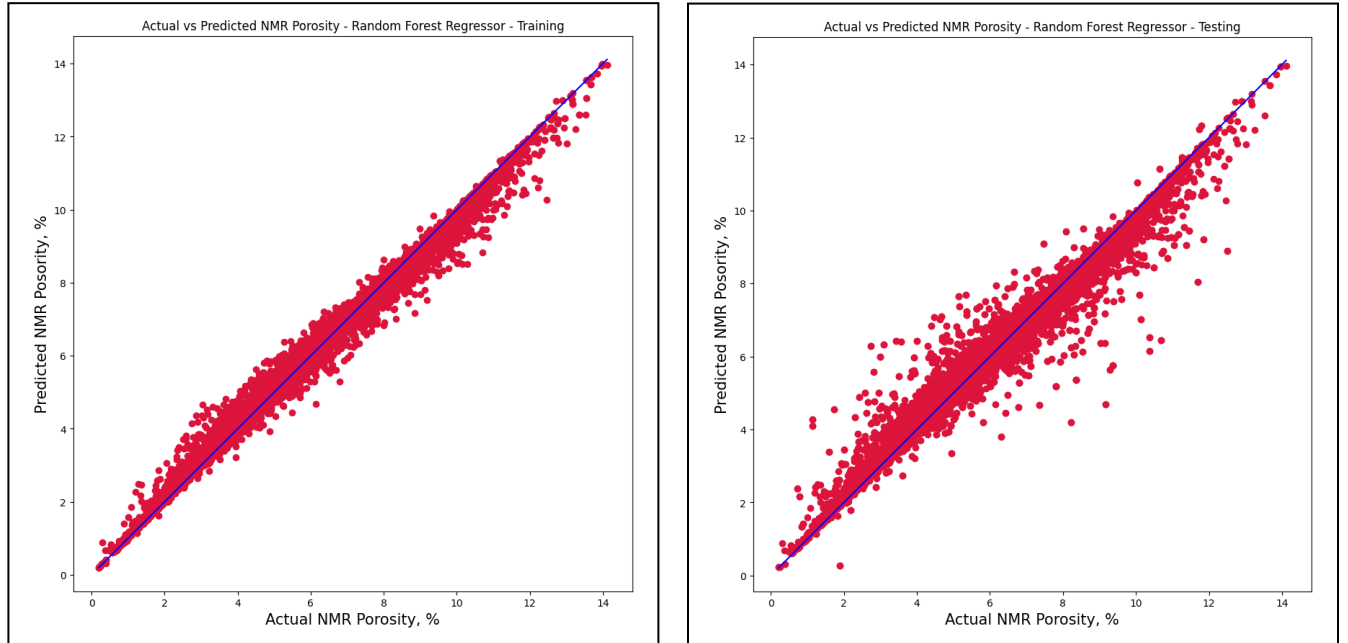


Figure 48: The correlation between the predicted values and the actual values of the total NMR porosity using the Random Forest Regressor algorithm for the testing group of the compiled data set.

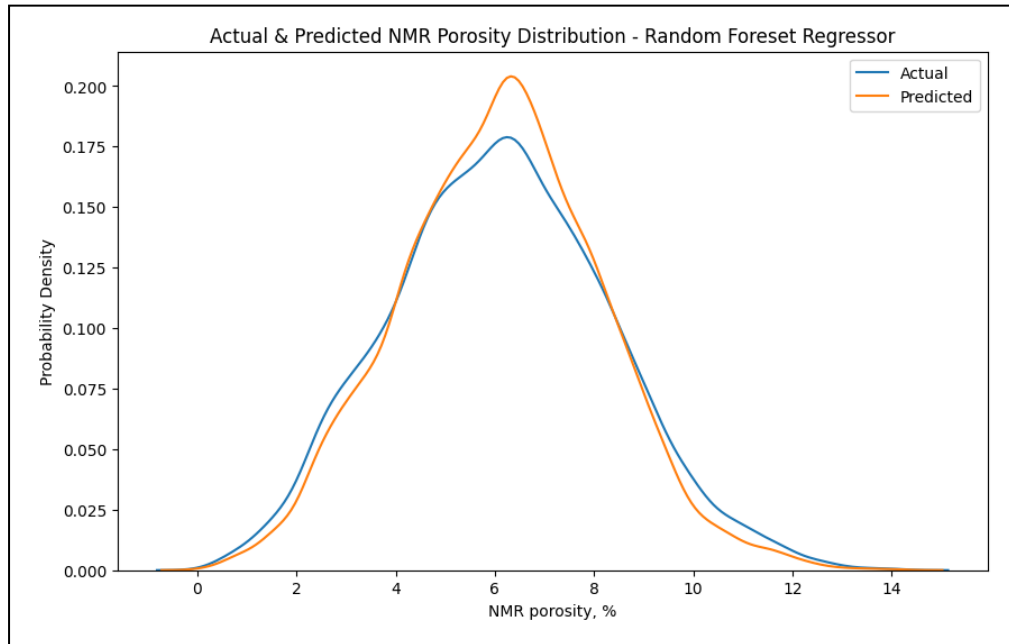


Figure 49: The distribution of the actual and the predicted values of the total NMR porosity using Random Forest Regressor algorithm for the testing group of the compiled data set.

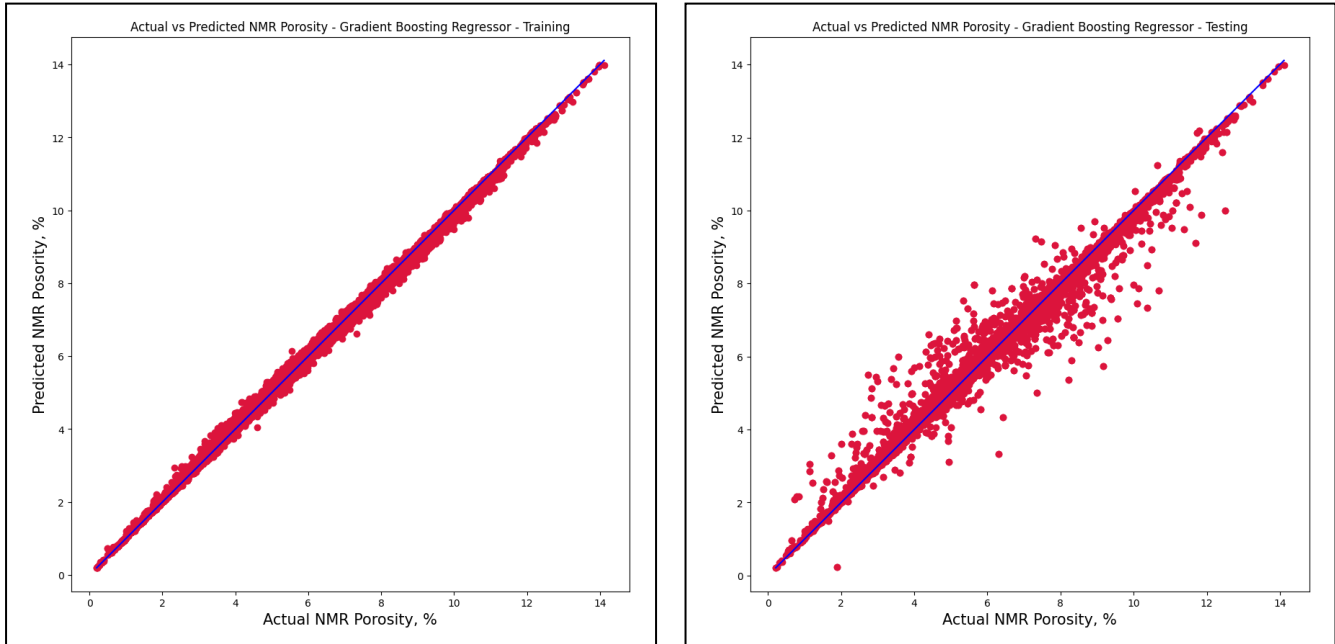


Figure 50: The correlation between the predicted values and the actual values of the total NMR porosity using the Gradient Boosting Regressor algorithm for the testing group of the compiled data set.

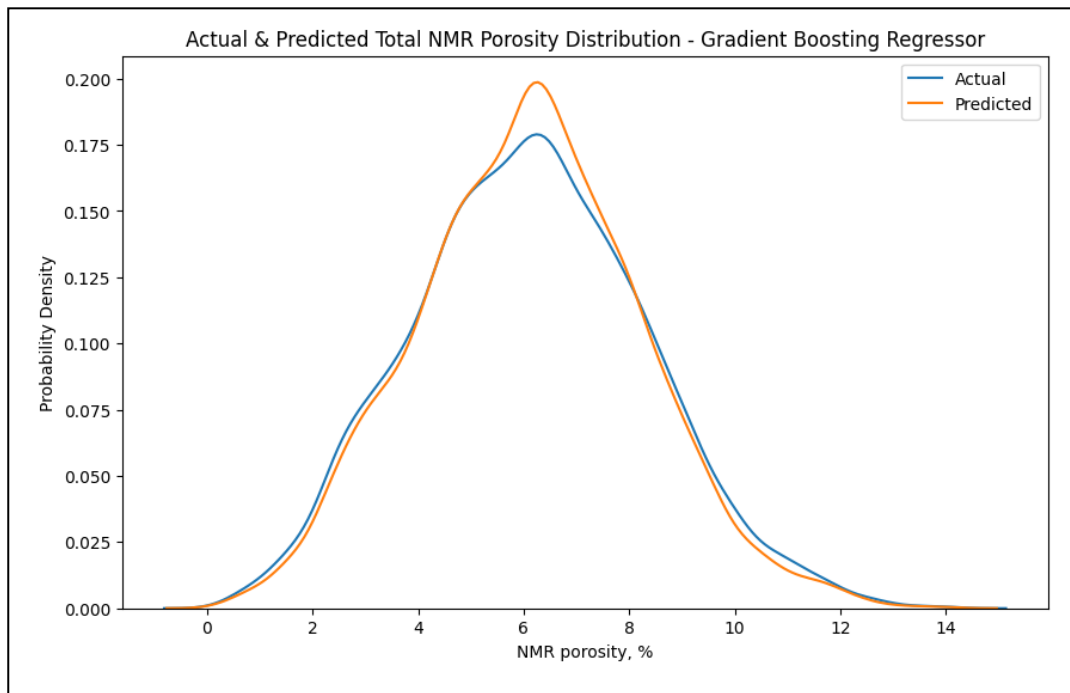


Figure 51: The distribution of the actual and the predicted values of the total NMR porosity using Gradient Boosting Regressor algorithm for the testing group of the compiled data set.

Remarks:

- From the scatter charts (figure-42, 44, 46, and 48), the predicted values and the actual values of the total NMR porosity are highly correlated. This indicates the highly matching for the prediction results from both Random Forest Regressor and Gradient Boosting Regressor algorithms.
- Figures (43, 45, 47, and 49) are displaying the distribution of the actual and the predicted values of total NMR porosity for the testing group. Overall, the two lines are following the same distribution, except for the porosity measurements between 5% and 8%, where the predicated values are having higher occurrence. This might lead to overestimating the total NMR porosity at that range.

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 Interpretation of the Findings

The interpretations of the results for predicting total NMR porosity using machine learning algorithms are summarized below.

- Three sets on data that includes several conventional logs measurements were used as input data for a machine learning model.
- Two of the three sets have the same log measurements and one has additional well logs including detailed log measurement and lithology measurements.
- The lithology measurements from data set number 2 were used to confirm that used data belongs to carbonate rocks.
- The three data sets are data set number 1 (6 inputs and 1 output with 13933 data points), data set number 2 (24 input and 1 output with 5802 data points), and the compiled data set (6 inputs and 1 output with 19735 data points).
- A statistical analysis was conducted before starting the model training to evaluate the relationship between the input variables of each data set with the output variable (total NMR porosity).
- The three data sets have similar range of NMR porosity measurements; data set number 1 (0.20 %- 14.11 %), data set number 2 (0.21% – 13.99%), and the compiled data set (0.20% – 14.11 %).

- The statistical analysis indicated the relationship between the input variables and the total NMR porosity as they are either positively or negatively correlated with the total NMR porosity.
- The three data sets were divided into 70% of the data for training group and 30% for the testing group.
- The three data sets were trained and tested using 10 different regression algorithms in 4 main machine learning models.
- The results of each algorithm were evaluated using 3 error metrics; correlation coefficient (R^2), absolute average percentage error (AAPE), and root mean square error (RMSE).
- Several steps of model enhancement have been applied on the three data sets that include tuning of the algorithms hyperparameters, clustering the data points, and re-tuning of the algorithms hyperparameters after including the data clusters. Figure-50 presents a simplified workflow for the applied enhancements and models' adjustments to deliver acceptable results of machine learning models.
- Five (5) clustering algorithm were computed and evaluated using Silhouette Coefficient to divide the data sets into several number of clusters.
- Five (5) out of the ten (10) algorithms did not deliver any major improvements in the accuracy of the total NMR prediction. These algorithms are Linear Regression, Decision Tree Regressor, XGBRFRegressor, LGBMRegressor, and keras (ANN). Some of these algorithms require further development in order to accommodate the number of features (inputs) and data points of the three data set. Other algorithms might not be able to suitable for this kind of problem such as Linear Regression.

- Three of the algorithms, which are XGBRegressor, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) Regressor, have demonstrated good results with the training group, but not with the testing group. Further analysis might be required to address the objective of achieving the requires results.
- The machine learning algorithms Random Forest Regressor and Gradient Boosting Regressor have managed to deliver satisfactory results for the prediction of total NMR porosity using data set number 1 and the complied data set.
- For data set number 1:
 - The results of the Random Forest Regressor algorithm are 100% and 91% for the correlation coefficient, $1.01e-5$ and 0.2842 for the AAPE, 0.0012 and 0.6782 for the RMSE for both of the training and the testing groups, respectively.
 - The Gradient Boosting Regressor results are 99% and 93% for the correlation coefficient, 0.1966 and 0.3960 for the AAPE, 0.2587 and 0.5771 for the RMSE for both of the training and the testing groups, respectively.
- For the compiled data set:
 - The results of the Random Forest Regressor algorithm are 99% and 95% for the correlation coefficient, 0.0948 and 0.2607 for the AAPE, 0.1716 and 0.4776 for the RMSE for both of the training and the testing groups, respectively.
 - The Gradient Boosting Regressor results are 100% and 97% for the correlation coefficient, 0.1157 and 0.2177 for the AAPE, 0.1568 and 0.3765 for the RMSE for both of the training and the testing groups, respectively.

- The cross validation of five (5) randomly selected folds during the training of the machine learning model for the two algorithms Random Forest Regressor and Gradient Boosting Regressor have delivered the score of 86% and 90% for data set number 1, respectively, 92% and 96% for the complied data set.
- The results of the machine learning algorithms using data set number 2 did not provide acceptable results comparing to the other data sets. The reason for issue could be the high number of features as the developed models might not be able to accommodate all the features. Additional data preparation and analysis before re-testing the model might enhance the model results.
- The additional number of data points for the complied data set have provided higher accuracy results for the machine learning model. This highlights the importance of availing sufficient amount of data for the machine learning solutions.

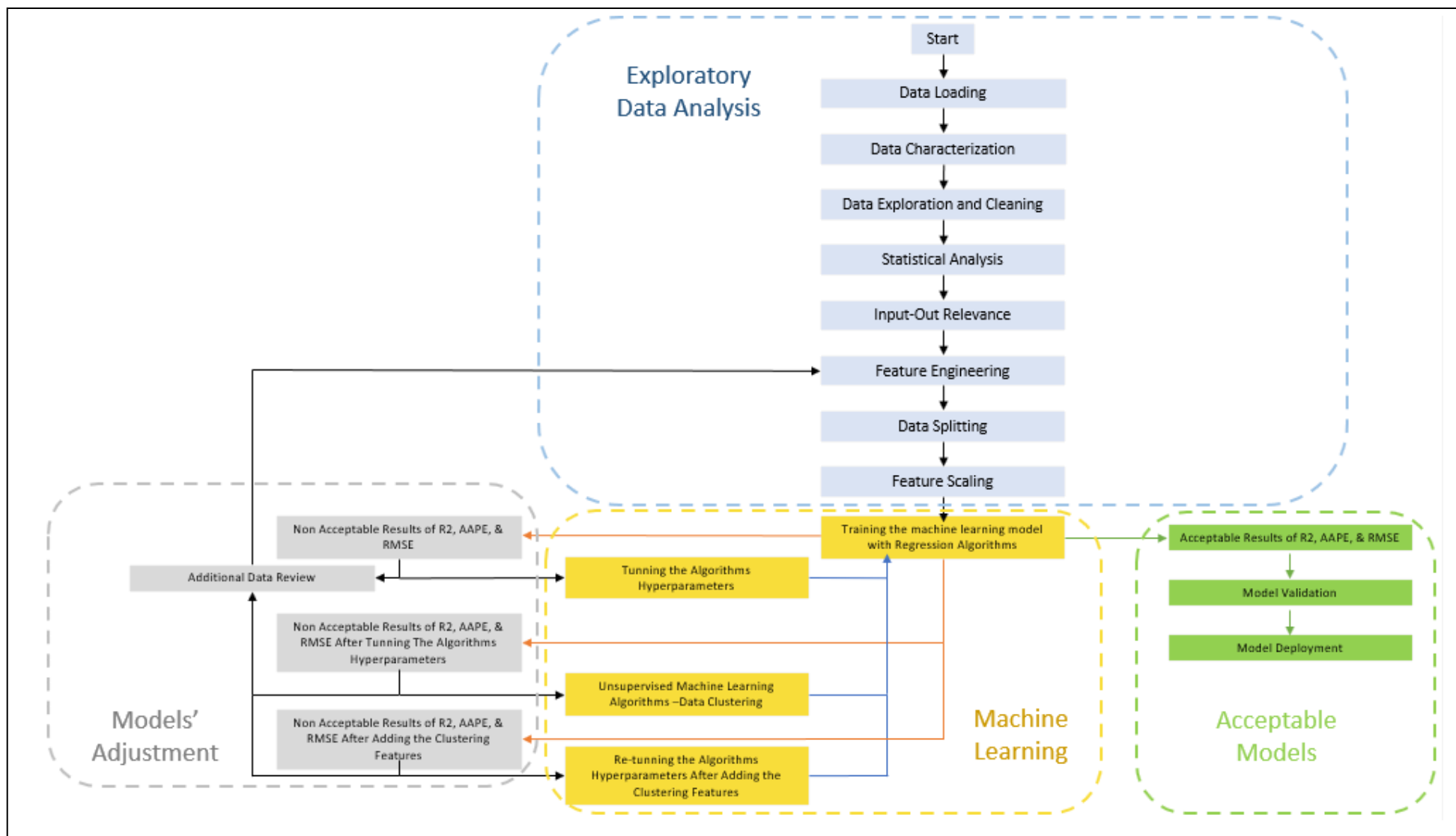


Figure 52: A simplified workflow for the used step to predict the total NMR porosity using machine learning.

6.2 Limitations of the Study

Using the previous machine learning methods to predict porosity has limitation. As the machine learning models were established and trained by using the results from the NMR porosities, the similar limitation that impact the NMR reading will affect the porosity prediction from the models. For this reasons, the associated reservoir conditions need to be considered during the process of porosity prediction by using machine learning approach.

6.2.1 Shale Impact on NMR Signals

The prediction of porosity from the NMR tool depends on the measurement of the longitudinal (T1) and transverse (T2) relaxation times. The measurement of the relaxation times is different in the porous media than in a formation with confined fluids, where the measurement will be considering the interactions with the fluid's hydrogen atoms only. As for the porous media, the interactions will be counted for the hydrogen atoms in the fluids and the confining surfaces. In addition, the diffusion impact will make a difference in the measurement in the porous media as it will act as semi-permeable barrier. This impact will increase significantly in a formation with different pore sizes such as in shaly formations, where both micro-pores and nano-pores exist. Besides that, the confined fluids within the clay layers will lead to overestimate the porosity measurements from NMR tool [90].

6.2.2 NMR Analysis Discrepancy from LWD

There are certain circumstances where the NMR tool might provide data with low quality. One way to maintain the data quality is by maintaining the running speed for the NMR tool. The objective of maintaining the running speed is to reduce the mechanical motion impact on the tool to provide good measurements for the relaxation times (T1) and (T2). Running the NMR tool during the drilling operation, which is knowing as Logging While

Drilling (LWD), might not provide the suitable conditions to run the NMR tool. This can be observed in the drilling operations where high penetration rates and running the drilling bits at high speed are required. These conditions will not provide the minimum required time for the NMR tool to measure the relaxation times accurately. In addition, the resulted vibration from the drilling process will lead to incorrect reading for (T1) and (T2). Moreover, the NMR tool needs to be allocated in the center of the borehole to deliver accurate readings for the surrounding formation data, which could not be granted during the drilling operations [33].

6.2.3 NMR Analysis in Gas Hydrate-Bearing Sections

In some situation, the prediction of porosity by using NMR tools might be underestimated. This can be encountered in the sections of gas hydrate-bearing, which is consisting from water and gas to create crystalline solids. The interaction between the molecules of gas hydrate might not be detected by the NMR tool signal. Even though, there are many laboratory studies that have been analyzing the gas hydrate sample to determine the nuclear magnetic properties, but still there is no accurate research that can provide the expected reaction of gas hydrate molecules in the wellbore. Furthermore, some analysis of gas hydrate clathrates response with NMR signals provides the same response as some rock matrices. This means that the measurement of porosity in some sections might be less than the actual porosity [33].

6.2.4 Limitation of Machine Learning Models

Although the concept of using machine learning to predict total porosity might be alternative and sufficient solution in many cases, it might encounter challenges that impact

the prediction efficiency. As the model was developed based on historical and limited sets of data with certain reservoir conditions, where the range of the NMR porosity varies between 0.20% to 14.11%, introducing new data points that represent different conditions could reduce the model's accuracy. This will increase the necessity for more computing resources that handle the complexity of adding new data with new conditions. For this, a continuous monitoring and update for the model is required to maintain the acceptable accuracy level [8].

6.3 Conclusion

The prediction of porosity in carbonate reservoir is a major challenge due to the high complexity of the carbonate rocks. The Nuclear Magnetic Resonance (NMR) well logs can provide high prediction of the total porosity in carbonate reservoirs over other traditional methods. The availability of NMR logs is limited and cannot be utilized in most cases of porosity prediction. The prediction of total NMR porosity using machine learning methods can deliver high accuracy measurements and reduce the need of traditional porosity measurement methods. Below is the conclusion of this research:

- The total porosity of the Nuclear Magnetic Resonance (NMR) well logs in carbonate reservoirs is predictable using machine learning algorithms.
- The conventional well logs including gamma ray, caliper, neutron porosity, density, photoelectric can be utilized as alternative source for total NMR porosity instead of running the NMR well log.
- The two machine learning algorithms Random Forest Regressor and Gradient Boosting Regressor were able to predict total NMR porosity with high accuracy.
- The machine learning model using the two algorithms have been evaluated using

the correlation coefficient (R^2), average absolute percentage error (AAPE), and root mean squared error (RMSE) for training, testing, and validation groups of data and it showing high correlation and low error for the three groups.

- This initial solution of machine learning can provide an alternative solution for measuring total porosity in carbonate by traditional measurement methods.
- The research has led to generate a simplified workflow for predicting total NMR porosity in carbonate reservoir using the machine learning algorithms. The workflow illustrates the general actions to handle different situations of data preparations, model configurations, and results validation.
- The generated workflow can be utilized for similar problems of predicting petrophysical properties using machine learning techniques.
- The outcomes of the research support the utilization of machine learning algorithms specifically and artificial intelligence in general to handle challenging issues in the petroleum industry.

References

- [1] “3.2.2: Porosity from Well Logs | PNG 301: Introduction to Petroleum and Natural Gas Engineering.” Wwww.e-Education.psu.edu, www.e-education.psu.edu/png301/node/835#:~:text=The%20most%20common%20method%20of. Accessed 30 Dec. 2022.
- [2] “Correlation Coefficient - Definition, Formula, Properties, Examples.” BYJUS, byjus.com/jee/correlation-coefficient/#:~:text=The%20covariance%20of%20two%20variables.
- [3] “Exploring the Impact of Artificial Intelligence on Offshore Oil and Gas.” Offshore Technology, 15 May 2019, www.offshore-technology.com/features/application-of-artificial-intelligence-in-oil-and-gas-industry/#:~:text=The%20AI%20intuitively%20links%20information. Accessed 18 Dec. 2022.
- [4] “What Is Model Tuning for ML Models? | Domino Data Lab.” Wwww.dominodatalab.com, www.dominodatalab.com/data-science-dictionary/model-tuning#:~:text=Model%20tuning%20is%20the%20experimental.
- [5] Abbasi, J., Zhao, J., Ahmed, S., Jiao, L., Andersen, P. S., & Cai, J. (2022, October 15). Prediction of permeability of tight sandstones from mercury injection capillary pressure tests assisted by a machine-learning approach. *Capillarity*, 5(5), 91–104. <https://doi.org/10.46690/capi.2022.05.02>
- [6] Adelman, Melanie Milner 1988-. “Building Static Geological Models in the Presence of Missing Values and Sparse Data.” Uh-Ir.tdl.org, 1 Dec. 2019, uh-ir.tdl.org/handle/10657/5544. Accessed 18 Dec. 2022.
- [7] Ahmed S., Abdulmalek, Mahmoud, Ahmed Abdulhamid, Elkatatny, Salaheldin, Mahmoud, Mohamed, and Abdulazez Abdulraheem. "Prediction of Pore and Fracture Pressures Using Support Vector Machine." Paper presented at the International Petroleum Technology Conference, Beijing, China, March 2019. doi: <https://doi.org/10.2523/IPTC-19523-MS>
- [8] Alangari, N., El Bachir Menai, M., Mathkour, H., & Almosallam, I. (2023, August 21). Exploring Evaluation Methods for Interpretable Machine Learning: A Survey. *Information*, 14(8), 469. <https://doi.org/10.3390/info14080469>
- [9] Alarfaj, M. K., Abdulraheem, A.. , Al-Majed, A.. , and E.. Hossain. "Estimating Layers Deliverability in Multi-Layered Gas Reservoirs Using Artificial Intelligence." Paper presented at the SPE Saudi Arabia Section Technical Symposium and Exhibition, Al-Khobar, Saudi Arabia, April 2012. doi: <https://doi.org/10.2118/160893-MS>
- [10] Alkhidir, Khalid Elyas Mohamed Elameen & Benzagouta, Mohammed & Аль-Хидир, К & Бензагута, М & Колледж, Инженерный & Сауда, Университет. (2013). TIGHT CARBONATE RESERVOIR CHARACTERIZATION

ХАРАКТЕРИСТИКА ПЛОТНОГО КАРБОНАТНОГО КОЛЛЕКТОРА. 206 - 217.

- [11] Alnuaimi, M. M. 2018. Using Artificial Intelligence and Machine Learning to Develop Synthetic Well Logs. Graduate Thesis, Dissertations, and Problem Reports. 3747. DOI: <https://doi.org/10.33915/etd.3747>
- [12] Alpaydin, E. (2014, August 22). Introduction to Machine Learning. MIT Press. http://books.google.ie/books?id=NP5bBAAAQBAJ&printsec=frontcover&dq=Introduction+to+Machine+Learning&hl=&cd=5&source=gbs_api
- [13] Andersen, Pål & Skjeldal, Miranda & Augustsson, Carita. (2022). Machine Learning Based Prediction of Porosity and Water Saturation from Varg Field Reservoir Well Logs. 10.2118/209659-MS.
- [14] Andersen, Pål Østebø, Skjeldal, Miranda, and Carita Augustsson. "Machine Learning Based Prediction of Porosity and Water Saturation from Varg Field Reservoir Well Logs." Paper presented at the SPE Europec - Europe Energy Conference featured at the 83rd EAGE Annual Conference & Exhibition, Madrid, Spain, June 2022. doi: <https://doi.org/10.2118/209659-MS>
- [15] Andersen, Pål Østebø, Skjeldal, Miranda, and Carita Augustsson. "Machine Learning Based Prediction of Porosity and Water Saturation from Varg Field Reservoir Well Logs." Paper presented at the SPE Europec - Europe Energy Conference featured at the 83rd EAGE Annual Conference & Exhibition, Madrid, Spain, June 2022. doi: <https://doi.org/10.2118/209659-MS>
- [16] Andreola, Fernanda & Leonelli, Cristina & Romagnoli, Marcello & Miselli, Paola. (2000). Techniques Used to Determine Porosity. American Ceramic Society Bulletin. 79. 49-52.
- [17] Anifowose, Fatai , and Abdulazeez Abdulraheem. "Prediction of Porosity and Permeability of Oil and Gas Reservoirs Using Hybrid Computational Intelligence Models." Paper presented at the North Africa Technical Conference and Exhibition, Cairo, Egypt, February 2010. doi: <https://doi.org/10.2118/126649-MS>
- [18] Appel, Matthias. (2004). Nuclear magnetic resonance and formation porosity. Petrophysics. 45. 296-307.
- [19] Ayyaz Mustafa, Ayyaz, Zeeshan Tariq, Zeeshan, Mohamed Mahmoud, Mohamed, and Abdulazeez Abdulraheem. "A Data Driven Machine Learning Approach to Predict the Nuclear Magnetic Resonance Porosity of the Carbonate Reservoir." Paper presented at the International Petroleum Technology Conference, Riyadh, Saudi Arabia, February 2022. doi: <https://doi.org/10.2523/IPTC-22081-MS>
- [20] Ayyaz Mustafa, Ayyaz, Zeeshan Tariq, Zeeshan, Mohamed Mahmoud, Mohamed, and Abdulazeez Abdulraheem. "A Data Driven Machine Learning Approach to Predict the Nuclear Magnetic Resonance Porosity of the Carbonate Reservoir." Paper presented at the International Petroleum Technology Conference, Riyadh, Saudi Arabia, February 2022. doi: <https://doi.org/10.2523/IPTC-22081-MS>
- [21] Berger, David L. (David Lemuel). Lithologic Properties of Carbonate-Rock Aquifers at Five Test Wells in the Coyote Spring Valley Area, Southern Nevada, as Determined from Geophysical Logs / by David L. Berger ; Prepared in

- Cooperation with the State of Nevada, the Las Vegas Valley Water District, and the U.S. Bureau of Reclamation. Carson City, Nev: U.S. Dept. of the Interior, U.S. Geological Survey, 1992. Print.
- [22] Bishop, C. M. (2006, August 17). Pattern Recognition and Machine Learning. Springer Verlag.
http://books.google.ie/books?id=qWPwnQEACAAJ&dq=Pattern+Recognition+and+Machine+Learning&hl=&cd=2&source=gbs_api
 - [23] Bontempi. (n.d.). Statistical foundations of machine learning (INFO-F-422).
<https://ai.vub.ac.be/sites/default/files/4.%20Hypothesis%20Testing.pdf>
 - [24] BREIMAN. (2001). Random Forests. Kluwer Academic Publishers, 45, 5–32.
 - [25] Bruna, Pierre-Olivier. (2013). Tight Carbonates Reservoir Properties Evolution: A case study of Lower Cretaceous hemipelagic limestone from The Northern Provence Region. 10.13140/RG.2.1.3463.8247.
 - [26] Bruna, Pierre-Olivier. (2013). Tight Carbonates Reservoir Properties Evolution: A case study of Lower Cretaceous hemipelagic limestone from The Northern Provence Region. 10.13140/RG.2.1.3463.8247.
 - [27] Burchette, Trevor P. “Carbonate Rocks and Petroleum Reservoirs: A Geological Perspective from the Industry.” Geological Society, London, Special Publications, vol. 370, no. 1, 2012, pp. 17–37, 10.1144/sp370.14
 - [28] Burchette, Trevor. (2012). Carbonate rocks and petroleum reservoirs: A geological perspective from the industry. Geological Society of London Special Publications. 370. 17-37. 10.1144/SP370.14.
 - [29] Burns, Ed, et al. “What Is Artificial Intelligence (AI)?” SearchEnterpriseAI, TechTarget, Feb. 2022, www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.
 - [30] Chang, Dahai, Vinegar, Harold J., Morriss, Chris, and Chris Straley. "Effective Porosity, Producing Fluid And Permeability In Carbonates From Nmr Logging." Paper presented at the SPWLA 35th Annual Logging Symposium, Tulsa, Oklahoma, June 1994.
 - [31] Chelliah, I. (2022, January 2). Correlation Coefficient—Clearly Explained. All About AI-ML. <https://indhumathychelliah.com/2020/12/16/correlation-coefficient%E2%80%8A-%E2%80%8Aclearly-explained/#:~:text=In%20machine%20learning%2C%20before%20we,as%20a%20measure%20of%20association>.
 - [32] Choquette, P. W., and L. C. Pray, 1970, Geological nomenclature and classification of porosity in sedimentary carbonates: AAPG Bulletin, v. 54, p. 207-250
 - [33] Collett, T. & Lee, Myung & Goldberg, David & Rack, Frank & Williams, Trevor. (2006). 27. Data Report: Nuclear Magnetic Resonance Logging while Drilling, ODP Leg 204. Proc. ODP, Sci. Results. 204. 1-22. 10.2973/odp.proc.sr.204.106.2006.
 - [34] Comaniciu, D., & Meer, P. (2002, May). Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), 603–619. <https://doi.org/10.1109/34.1000236>

- [35] Copeland, B.J.. "artificial intelligence". Encyclopedia Britannica, 11 Nov. 2022, <https://www.britannica.com/technology/artificial-intelligence>. Accessed 17 December 2022.
- [36] Correlation and Regression with R. (n.d.). https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression_print.html
- [37] Das, Troyee & Mukherjee, Soumyajit. (2020). Porosity in Carbonates. 10.1007/978-3-030-13442-6_2.
- [38] Dautriat, Jérémie, et al. "Hydromechanical Behavior of Heterogeneous Carbonate Rock under Proportional Triaxial Loadings." *Journal of Geophysical Research*, vol. 116, no. B1, 26 Jan. 2011, 10.1029/2009jb000830. Accessed 2 Nov. 2022.
- [39] Dlubac, Katherine & Knight, Rosemary & Song, Yi-Qiao & Bachman, H. & Grau, Ben & Cannia, Jim & Williams, John. (2013). Use of NMR logging to obtain estimates of hydraulic conductivity in the High Plains aquifer, Nebraska, USA. *Water Resources Research*. 49. 1871-1886. 10.1002/wrcr.20151.
- [40] El-Samak, A. F., & Ashour, W. (2015, October 1). Optimization of Traveling Salesman Problem Using Affinity Propagation Clustering and Genetic Algorithm. *Journal of Artificial Intelligence and Soft Computing Research*, 5(4), 239–245. <https://doi.org/10.1515/jaiscr-2015-0032>
- [41] Ester, Kriegel, Sander, & Xu. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Association for the Advancement of Artificial Intelligence (AAAI)*, 226–229. <https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf>
- [42] Ezekwe N. *Petroleum Reservoir Engineering Practice*. Upper Saddle River NJ: Prentice Hall; 2011.
- [43] F. Jerry Lucia; *Rock-Fabric/Petrophysical Classification of Carbonate Pore Space for Reservoir Characterization*. AAPG Bulletin 1995;; 79 (9): 1275–1300. doi: <https://doi.org/10.1306/7834D4A4-1721-11D7-8645000102C1865D>
- [44] Farmanov, Ravan, Feldmann, Felix, Mathew, Eric Sonny, Tembely, Moussa, Al-Shalabi, Emad W., AlAmeri, Waleed, Masalmeh, Shehadeh, and Ali AlSumaiti. "Application of Machine Learning for Estimating Petrophysical Properties of Carbonate Rocks Using NMR Core Measurements." Paper presented at the SPE Reservoir Characterisation and Simulation Conference and Exhibition, Abu Dhabi, UAE, January 2023. doi: <https://doi.org/10.2118/212625-MS>
- [45] Frey, B. J., & Dueck, D. (2007, February 16). Clustering by Passing Messages Between Data Points. *Science*, 315(5814), 972–976. <https://doi.org/10.1126/science.1136800>
- [46] Friedman, J. H. (2001, October 1). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5). <https://doi.org/10.1214/aos/1013203451>
- [47] G. (2024, April 29). Regression using LightGBM. *GeeksforGeeks*. <https://www.geeksforgeeks.org/regression-using-lightgbm/>

- [48] G. (2024, April 4). Introduction to XGBoost Algorithm in Machine Learning. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/>
- [49] GBM in Machine Learning - Javatpoint. (n.d.). [www.javatpoint.com. https://www.javatpoint.com/gbm-in-machine-learning](https://www.javatpoint.com/gbm-in-machine-learning)
- [50] Georgi, Daniel T., Shorey, David S., and Galy M. Ostroff. "Integration Of Nmr And Conventional Log Data For Improved Petrophysical Evaluation Of Shaly Sands." Paper presented at the SPWLA 40th Annual Logging Symposium, Oslo, Norway, May 1999.
- [51] Géron, A. (2019, September 5). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. "O'Reilly Media, Inc." [http://books.google.ie/books?id=HnetDwAAQBAJ&printsec=frontcover&dq=Au relien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligen&hl=&cd=1&source=gbs_api](http://books.google.ie/books?id=HnetDwAAQBAJ&printsec=frontcover&dq=Au+relien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligen&hl=&cd=1&source=gbs_api)
- [52] Glover, Paul. "Formation Evaluation MSc Course Notes." [Www.academia.edu, www.academia.edu/25574527/Formation_Evaluation_MSc_Course_Notes](http://www.academia.edu/25574527/Formation_Evaluation_MSc_Course_Notes).
- [53] Glover, Paul. "Petrophysics MSc Course Notes." [Www.academia.edu, www.academia.edu/11160139/Petrophysics_MSc_Course_Notes](http://www.academia.edu/11160139/Petrophysics_MSc_Course_Notes). Accessed 18 Dec. 2022.
- [54] Guo, Boyun. "Petroleum Reservoir Properties." Well Productivity Handbook, 2019, pp. 17–51, 10.1016/b978-0-12-818264-2.00002-6.
- [55] H. (2023, September 5). Understanding Gradient Boosting - Hemashreekilari - Medium. [Medium. https://medium.com/@hemashreekilari9/understanding-gradient-boosting-632939b98764](https://medium.com/@hemashreekilari9/understanding-gradient-boosting-632939b98764)
- [56] Han, J., Kamber, M., & Pei, J. (2011, June 9). Data Mining: Concepts and Techniques. Elsevier. [http://books.google.ie/books?id=pQws07tdpjoC&printsec=frontcover&dq=Data+ Mining%0D%0AConcepts+and+Techniques&hl=&cd=1&source=gbs_api](http://books.google.ie/books?id=pQws07tdpjoC&printsec=frontcover&dq=Data+Mining%0D%0AConcepts+and+Techniques&hl=&cd=1&source=gbs_api)
- [57] Helvacı, Cahit. (2021). Borates: Encyclopedia of Geology, 2nd Edition, Vol.1.
- [58] Hiatt, Eric E., and Kyser Kurt. "CHAPTER 4. LINKS between DEPOSITIONAL and DIAGENETIC PROCESSES in BASIN ANALYSIS: POROSITY and PERMEABILITY EVOLUTION in SEDIMENTARY ROCKS." Academia, 2017,
- [59] Hou, Jue & Zhao, Lun & Zeng, Xing & Zhao, Wenqi & Chen, Yefei & Li, Jianxin & Wang, Shuqin & Wang, Jincai & Song, Heng. (2022). Characterization and Evaluation of Carbonate Reservoir Pore Structure Based on Machine Learning. Energies. 15. 7126. 10.3390/en15197126.
- [60] Jain, S. (2024b, May 29). An Overview of Regularization in Deep Learning (with Python code). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques>

- [61] Kumar, Arvind. (2012). Artificial Neural Network as a Tool for Reservoir Characterization and its Application in the Petroleum Engineering. 10.4043/22967-MS.
- [62] Kumar, Sunil & Chong, Ilyoung. (2018). Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. International Journal of Environmental Research and Public Health. 15. 2907. 10.3390/ijerph15122907.
- [63] Lupinacci, Wagner & Gomes, Livia & Ferreira, Danilo & Bijani, Rodrigo & Freire, Fernando. (2020). An integrated approach for carbonate reservoir characterization: a case study from the Linguado Field, Campos Basin. Brazilian Journal of Geology. 50. 1-12. 10.1590/2317-4889202020190103.
- [64] M. Ahr, Wayne. "A New Genetic Classification of Carbonate Porosity and Its Application to Reservoir Characterization, by Wayne M. Ahr, #40357 (2008)." Wwww.searchanddiscovery.com, 24 Oct. 2008, www.searchanddiscovery.com/documents/2008/08177ahr/. Accessed 18 Dec. 2022.
- [65] Mackenzie, Fred T, and John W Morse. "Sedimentary Carbonates through Phanerozoic Time." *Geochimica et Cosmochimica Acta*, vol. 56, no. 8, Aug. 1992, pp. 3281–3295, 10.1016/0016-7037(92)90305-3. Accessed 14 Nov. 2022.
- [66] Mackenzie, W S, et al. *A Colour Atlas of Rocks and Minerals in Thin Section*. Boca Raton, Crc Press, 2015.
- [67] McKinney, W. (2017, September 25). Python for Data Analysis. "O'Reilly Media, Inc." http://books.google.ie/books?id=UiM3DwAAQBAJ&printsec=frontcover&dq=Python+for+Data+Analysis&hl=&cd=1&source=gb_api
- [68] Miller, M.N., Paltiel, Z., Gillen, M.E., Granot, J., and J.C. Bouton. "Spin Echo Magnetic Resonance Logging: Porosity and Free Fluid Index Determination." Paper presented at the SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, September 1990. doi: <https://doi.org/10.2118/20561-MS>
- [69] Mohammad Hossain, Touhid & Watada, Junzo & Jian, Zhiwen & Sakai, Hiroshi & a Rahaman, Md & Aziz, Associate Professor Dr Izzatdin. (2020). MISSING WELL LOG DATA HANDLING IN COMPLEX LITHOLOGY PREDICTION: AN NIS APRIORI ALGORITHM APPROACH. International journal of innovative computing, information & control: IJICIC. 16. 1077-1091. 10.24507/ijicic.16.03.1077.
- [70] Müller-Petke, Mike & Hiller, Thomas & Herrmann, Rolf & Yaramanci, Ugur. (2011). Reliability and limitations of surface NMR assessed by comparison to borehole NMR. Near surface geophysics. 9. 123-134. 10.3997/1873-0604.2010066.
- [71] Murphy, K. P. (2012, August 24). Machine Learning. MIT Press. http://books.google.ie/books?id=NZP6AQAAQBAJ&printsec=frontcover&dq=Machine+Learning+A+Probabilistic+Perspective+Kevin+P.+Murphy&hl=&cd=1&source=gb_api

- [72] Owolabi, Taoreed & Olatunji, Sunday. (2015). Feature Selection-Based ANN for Improved Characterization of Carbonate Reservoir. 10.2118/178029-MS.
- [73] Ozkaya, S. I.. "Using Probabilistic Decision Trees to Detect Fracture Corridors From Dynamic Data in Mature Oil Fields." SPE Res Eval & Eng 11 (2008): 1061–1070. doi: <https://doi.org/10.2118/105015-PA>
- [74] Popescu, Marius-Constantin & Balas, Valentina & Perescu-Popescu, Liliana & Mastorakis, Nikos. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 8.
- [75] Popescu, Marius-Constantin & Balas, Valentina & Perescu-Popescu, Liliana & Mastorakis, Nikos. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 8.
- [76] Prammer, M.G. "NMR Pore Size Distributions and Permeability at the Well Site." Paper presented at the SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, September 1994. doi: <https://doi.org/10.2118/28368-MS>
- [77] Quinlan, J.R. Induction of decision trees. Machine Learning 1, 81–106 (1986). <https://doi.org/10.1007/BF00116251>
- [78] Rafik, Baouche. (2016). Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. Egyptian Journal of Petroleum. journal homepage: www.elsevier.com/locate/petrol. 17. 10.1016/j.ejpe.2016.10.013.
- [79] Raschka, S., & Mirjalili, V. (2017, September 20). Python Machine Learning. Packt Publishing Ltd. http://books.google.ie/books?id=_plGDwAAQBAJ&printsec=frontcover&dq=Python+Machine+Learning+Second+Edition&hl=&cd=1&source=gbp_api
- [80] Rouhiainen, Lasse. "21 Artificial Intelligence Examples and Use Cases - Lasse Rouhiainen." Lasse Rouhiainen - International Keynote Speaker on Artificial Intelligence, Web3, Crypto, DAO, 5 Mar. 2021, lasserouhiainen.com/21-ai-examples/.
- [81] Rumelhart, David E. et al. "Learning representations by back-propagating errors." Nature 323 (1986): 533-536.
- [82] S. (2023, August 10). What is Statistical Analysis? Types, Methods and Examples. Simplilearn.com. <https://www.simplilearn.com/what-is-statistical-analysis-article#:~:text=Statistical%20analysis%20is%20a%20scientific,convert%20them%20into%20meaningful%20information.>
- [83] Secherla, Supriya. "Understanding Optimization Algorithms in Machine Learning." Medium, 15 July 2021, towardsdatascience.com/understanding-optimization-algorithms-in-machine-learning-edfdb4df766b.
- [84] Seilkhanov, Anuar , Otyunshin, Daniyar , Novopokrovskiy, Georgiy , and Emre Artun. "Comparison of Machine Learning Algorithms for the Development of a Forecasting Tool for Cyclic Gas Injection in Hydraulically-Fractured Wells." Paper presented at the SPE Annual Caspian Technical Conference, Virtual, October 2020. doi: <https://doi.org/10.2118/202507-MS>

- [85] Seltman. (2018, July 11). Experimental Design and Analysis. <https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>
- [86] Shaltami, Osama & Bustany, Ilas. (2021). Elemental geochemistry of sedimentary rocks – A review.
- [87] Stadtmuller, Marek. (2019). Well logging interpretation methodology for carbonate formation fracture system properties determination. *Acta Geophysica*. 67. 10.1007/s11600-019-00351-w.
- [88] Stephanie Glen. "Mean Absolute Percentage Error (MAPE)" From StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/mean-absolute-percentage-error-mape/>
- [89] Stephanie Glen. "RMSE: Root Mean Square Error" From StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/>
- [90] SWPU, Petroleum. (2016). Shale characteristics impact on Nuclear Magnetic Resonance (NMR) fluid typing methods and correlations. *Petroleum*. 2. 138-147.
- [91] T. (2023, May 4). Null Hypothesis: Learn definition, types, uses, and examples. Testbook. <https://testbook.com/maths/null-hypothesis#:~:text=The%20Formula%20for%20the%20Null,%E2%88%9AP%E2%88%92P0N.>
- [92] Tan, P. N., Steinbach, M., & Kumar, V. (2016, January 1). *Introduction to Data Mining*. Pearson Education India. http://books.google.ie/books?id=vIqqDwAAQBAJ&dq=Introduction+to+Data+Mining&hl=&cd=2&source=gbs_api
- [93] Tan, P.N. & Steinbach, Michael & Kumar, Vipin. (2005). *Cluster Analysis: Basic Concepts and Algorithms*. *Introduction to Data Mining*. 487-568.
- [94] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: an efficient data clustering method for very large databases. *SIGMOD Rec.* 25, 2 (June 1996), 103–114. <https://doi.org/10.1145/235968.233324>
- [95] Timur, A.. "Nuclear Magnetic Resonance Study Of Carbonate Rocks." *The Log Analyst* 13 (1972)
- [96] Timur, A.. "Producible Porosity And Permeability Of Sandstone Investigated Through Nuclear Magnetic Resonance Principles." *The Log Analyst* 10 (1969).
- [97] Tucker, Maurice E. *Sedimentary Petrology : An Introduction to the Origin of Sedimentary Rocks*. Oxford, Blackwell Science, 2001.
- [98] V. Kazmiruk (Ed.) *Scanning Electron Microscopy*. London, United Kingdom, IntechOpen, 2012 [Online]. Available from: <https://www.intechopen.com/books/1505> doi: 10.5772/1973
- [99] Vapnik, V. N. (2023, January 1). *The Nature of Statistical Learning Theory (Second Edition)*. [http://books.google.ie/books?id=z4PQzwEACAAJ&dq=The+Nature+of+Statistical+Learning+Theory+\(Second+Edition\)&hl=&cd=2&source=gbs_api](http://books.google.ie/books?id=z4PQzwEACAAJ&dq=The+Nature+of+Statistical+Learning+Theory+(Second+Edition)&hl=&cd=2&source=gbs_api)

- [100] Varhaug, Matt. THE DEFINING SERIES Basic Well Log Interpretation Matt Varhaug Senior Editor. 2016.
- [101] W.R. Moore, Y. Zee Ma, I. Pirie, Y. Zhang, Chapter 15 - Tight Gas Sandstone Reservoirs, Part 2: Petrophysical Analysis and Reservoir Modeling, Editor(s): Y. Zee Ma, Stephen A. Holditch, Unconventional Oil and Gas Resources Handbook, Gulf Professional Publishing, 2016, Pages 429-448, ISBN 9780128022382, <https://doi.org/10.1016/B978-0-12-802238-2.00015-8>.
- [102] Wade, C. (2022, June 3). XGBoost Alternative Base Learners - Towards Data Science. Medium. <https://towardsdatascience.com/xgboost-alternative-base-learners-a2dc72d97e64>
- [103] What Is a Decision Tree? | Master's in Data Science. (2024, May 8). CORP-MIDS1 (MDS). <https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/>
- [104] When using the K-Means Clustering Algorithm, is it possible to have a set of data which results in an Infinite Loop? (n.d.). Stack Overflow. <https://stackoverflow.com/questions/60312401/when-using-the-k-means-clustering-algorithm-is-it-possible-to-have-a-set-of-dat>
- [105] Wilke, C. O. (2019, March 18). Fundamentals of Data Visualization. O'Reilly Media. http://books.google.ie/books?id=XmmNDwAAQBAJ&printsec=frontcover&dq=Fundamentals+of+Data+Visualization&hl=&cd=1&source=gbs_api
- [106] Yehoshua, R. (2023, November 24). DBSCAN: Density-Based Clustering - Artificial Intelligence in Plain English. Medium. <https://ai.plainenglish.io/dbscan-density-based-clustering-aaebd76e2c8c> Yehoshua, R. (2023, November 24). DBSCAN: Density-Based Clustering - Artificial Intelligence in Plain English. Medium. <https://ai.plainenglish.io/dbscan-density-based-clustering-aaebd76e2c8c>

Vitae

Name	:Hussain Ali Alhamood
Nationality	:Saudi Arabia
Date of Birth	:11/30/1987
Email	:hussain_ah10@hotmail.com
Address	:8153 4B Alfirdwas, Sayhat 32434-4389, Eastern Province, Saudi Arabia
Academic Background	:MS of Science in Petroleum Engineering
Graduation:	June, 2024
Degrees:	Bachelor of Engineering in Petroleum Engineering University of Manchester, Manchester, UK June, 2010 Second Class Honor
Experience	Joined Saudi Aramco in Saudi Arabia in August, 2010 Worked as production engineer (2010 – 2011) Worked as drilling engineer (2012) Worked as production engineer (2013-2015) Worked as petroleum system analyst (2016 – present)

Publications

Altowilib, A.; AlSaihati, A.; Alhamood, H.; Alafnan, S.; Alarifi, S.

“Reserves Estimation for Coalbed Methane Reservoirs: A Review.” Sustainability 2020, 12, 10621.

<https://doi.org/10.3390/su122410621>

Alhamood, Hussain A., and Abdulazeez Abdulraheem.

"Permeability Prediction from Clay Volume Using Artificial Intelligence." Paper presented at the International Petroleum Technology Conference, Riyadh, Saudi Arabia, February 2022.

doi: <https://doi.org/10.2523/IPTC-22521-EA>

Alblushi, M.; Hassan, S.; Alhamood, H.; Zahdan, G.

“Business Process Mapping and Optimization in Oil and Gas”
MEOS 2023, 19. 213964.