



MISKOLCI
E G Y E T E M
UNIVERSITY OF MISKOLC



OTDK THESIS

EVALUATION OF SHALE RESERVOIRS BY ESTIMATING TOC, MINERAL
COMPOSITION AND POROSITY USING WELL LOGS

Prepared by:

Martins Uchenna Obidiegwu

NEPTUN CODE: ROGF8V

Supervised by:

Prof. Dr. Norbert Péter Szabó

30th December 2024

DECLARATION

Declaration of Authenticity

I, Martins Uchenna Obidiegwu, a student at the University of Miskolc, is aware of the criminal and disciplinary liability. I declare and confirm with my signature that I have prepared this thesis myself, that I have carried out the tests described in it and that I understand the presented results. I used data and information only from the sources listed in the bibliography. All parts that I have taken verbatim or paraphrased from another source have been clearly marked with the source.

Miskolc, date

 06/11/2024

Student's signature

Declaration of the Supervisor

I, the undersigned Professor Dr. Szabo Norbert Peter of University of Miskolc, Institute of Exploration Geosciences consider the TDK thesis suitable for submission.

Other comments, recommendation:

Miskolc, date 06.11.2024

Supervisor's signature



TABLE OF CONTENTS

TABLE OF CONTENTS	I
ABSTRACT	III
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Geological settings of the study area.....	3
1.3 Problem statement	5
1.4 Research objectives	6
1.5 Research Significance.....	7
1.6 General Methodology and Workflow	8
Chapter 2 Literature Review.....	9
2.1 Shale Gas as an Energy Source: Challenges and Advances	9
2.2 Mineralogical and Mechanical Properties of Shale Gas Reservoirs.....	9
2.3 Challenges in Mineralogical Evaluation.....	10
2.4 Data-Driven Approaches in Shale Reservoir Evaluation	10
2.5 Applications of AI in Shale Gas Development	10
2.6 Future Directions and Research Opportunities.....	12
2.7 Proposed Study: Formation G of the Ordos Basin	12
Chapter 3 Evaluation of Total Organic Carbon.....	14
3.1 Introduction	14
3.2 Methodology.....	14
3.2.1 The Passey Method	14
3.2.2 Multiple Linear Regression (MLR) Method	15
3.2.3 Machine Learning (ML) Method	15
3.2.4 Gradient Boost Method	16
3.2.5 Gradient Boost Prediction Model Setup	16
3.2.6 Adaptive Gradient Boost Method	17
3.2.7 Adaptive Gradient Boost Prediction Model Setup.....	17
3.2.8 Random Forest (RF) Method	17
3.2.9 Random Forest (RF) Prediction Model Setup.....	18
3.2.10 Stacking Technique	18

3.2.11 Architecture of Stacking.....	18
3.3 Results of TOC evaluation in Well 1.....	19
3.4 Performance and Comparative Analysis of All Methods for Well 1	26
3.5 Chapter Conclusion	26
Chapter 4 Evaluation of Mineral Composition and Porosity using Multiple Linear Regression (MLR) Method.....	28
4.1 Introduction	28
4.2 Methodology.....	28
4.3 Results	28
4.3.1 Mineral Composition Results from MLR Method in Well 1	28
4.3.2 Equation generated from the MLR method in Well 1.....	29
4.3.3 Analysis of Mineral Content and Porosity in Well 1 Using MLR Method....	31
4.4 Chapter Conclusions.....	32
Chapter 5 Evaluation of Mineral Composition and Porosity by Machine Learning Approach	33
5.1 Introduction	33
5.2 Methodology.....	33
5.3 Results	33
5.3.1 Gradient Boost Model Performance.....	35
5.3.2 Gradient Boost Results in Well 1	38
5.3.3 Extreme Gradient (XGBoost) Model Performance.....	40
5.3.4 XGBoost Result in Well 1.....	42
5.4 Performance and Comparative Analysis of MLR, Gradient Boost, and XG Boost in Evaluating Mineral Composition and Porosity.	44
5.5 Chapter Conclusions.....	44
Chapter 6 Conclusions and Recommendations	46
6.1 Conclusions	46
6.2 Future Research Prospectives	47
Appendix Own Results Summary	48
References	50
Acknowledgments	55

ABSTRACT

Formation G, a prominent part of the Mesozoic Yanchang Formation in the Ordos Basin, has attracted much attention because of its possible use as a reservoir for shale oil. The shale reservoir has excellent potential, but its TOC and mineral composition still need to be better understood and characterized, making it challenging to utilize and manage fully. This study seeks to estimate and assess the Total Organic Carbon (TOC) content using the Multiple Linear Regression, Passey, Machine Learning, and Stacking model. Multiple Linear Regression and Machine Learning (ML) were effectively utilized to estimate the mineralogical composition (clay, feldspar, quartz) and petrophysical features, such as the porosity of the Formation G shale.

The TOC results from Passey's method, MLR, and Stacking Model were in good agreement with the depth at which the TOC core data was measured, and it also shows a strong correlation with high gamma ray zones identified by gamma logs in the studied Shale Reservoir. Meanwhile, the MLR method outperformed the other methods, producing more reliable agreement with the core data for TOC and the evaluated TOC. Results for evaluating three critical minerals (clay, feldspar, and quartz) and porosity in the shale plays obtained from the MLR method indicated good predictive performance. However, the advanced machine learning techniques, namely Gradient Boost and Extreme Gradient Boost (XGBoost) models, demonstrated higher accuracy during both the training and testing phases, with error quantified using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the square of the correlation coefficient (R^2). The training phase of Gradient Boost models showed minimal errors and high correlation, whereas XGBoost models exhibited slightly lower accuracy metrics but demonstrated good predictive performance.

According to this research, the TOC could be analyzed using Passey's, MLR, ML, and Stacking Model as the results showed an acceptable agreement between the lab measurements of TOC and the predicted TOC. It was also established that MLR, ML, and Stacking models can be used to gain valuable insights into shale formations' TOC, mineral content, and porosity. This approach can optimize the evaluation and production of unconventional shale reservoirs worldwide. Additionally, this method may replace traditional techniques for determining TOC and mineralogy, providing a cost-effective and resource-efficient alternative, mainly when cores and mineralogy logs are unavailable for the Well. The methodologies and machine learning processes described in this study could potentially be applied to developing other shale gas and oil fields.

Chapter 1 Introduction

1.1 Background

Currently, the global energy pattern is dominated by the three traditional energy sources of oil, natural gas, and coal, supplemented by clean energy sources, such as nuclear, wind, and biomass energy, and vigorously developing new energy sources. With the proposal of a double-carbon target, the transformation of the global energy structure is gradually increasing (Zou, 2021). Natural gas is a clean, low-carbon, high-quality, and high-efficiency fossil fuel that can complement renewable energy. It is positioned as the most realistic choice for replacing coal, realizing cleanliness and low carbonization. It is the lifelong partner of renewable energy, playing a pillar role in transforming and developing energy from high to low carbon in the future (Bilgen & Sarıkaya, 2016; Lu & Zhao, 2013; Zou, 2021). Accelerating the development of the natural gas industry is one approach to consistently promoting the transformation of the energy consumption structure and building green, clean, safe, and efficient modern energy systems. It is also a viable option to strengthen energy conservation and emission reduction while actively responding to environmental issues, such as climate change.

Natural gases include both conventional and unconventional gases. Among them, unconventional natural gas is the main field of current and future exploration. It is a realistic and clean energy source with large-scale development and commercialization. Shale gas has great resource potential as an essential unconventional natural gas source. In recent years, through the establishment of China's shale gas national demonstration zone, a technical mode suitable for shale gas exploration and development has gradually formed through continuous exploration and breakthroughs. China has also become a third country after North America in achieving commercial development (Dong, 2016; S.-J. Wang, 2015; Xie, 2018). At the end of 2018, China had become the second-largest shale gas producer in the world (Jiao, 2019; W. Niu, 2023).

Shale oil is becoming an increasingly attractive area for exploration and development. China's vast shale oil deposits could be fully realized by improving drilling techniques and utilizing advanced technologies. However, it is important to note that shale, conventional, and gas reservoirs differ significantly in their physical characteristics, pore structures, and fluid occurrence modes. Oil and gas are self-generating and self-storing. However, industrial oil and gas flow can only be obtained via volume fracturing as there is no production capacity under natural conditions. The conventional reservoir depends on the physical qualities of the reservoir, such as porosity, permeability, oil saturation, and other variables.

Evaluation of shale reservoirs includes log characteristics, log R, and log curve overlap. Passey et al. (Q. R. Passey, 1990) proposed the most widely used log R approach, initially employed to identify significantly organic-rich source rocks. Shale has considerable thickness, good continuity, and slight variation in mineral content. Therefore, results obtained from the $\Delta\log R$ method were more promising in identifying high-quality shales.

The stress-strain curve for rocks is an intuitive and effective method of determining the qualitative brittleness of rocks. Thus, there are two primary methods for determining rock brittleness: the first is based on Young's modulus and Poisson's ratio approach, which does not consider the effect of stress and mineral content on the brittleness index as this approach is satisfactory for strata with wide variations in burial depth and mineral type. The second approach depends on the number of brittle minerals in the rock (e.g., feldspar, quartz, and carbonate). This method primarily considers the effect of a single mineral composition factor on rock brittleness and does not consider the impact of rock diagenesis and evolution on brittleness. When employing the mineral approach, it is necessary to identify the main types of brittle minerals and the diagenesis background in the study area and create a suitable evaluation model. Shale reservoirs should be evaluated based on both geological properties and engineering improvements. Numerous investigations have been conducted on shale oil geological properties, formation mechanisms, enrichment laws, volume transformations, and increased oil recovery (Zou, 2013). However, most shale evaluations have focused on a single element. Shale reservoirs' geological and engineering properties have rarely been combined to assess their sweet spots thoroughly. Research and development of shale oil formations have proven that even geological hotspots with high hydrocarbon generation potential are only sometimes engineering hotspots for volumetric fracturing. Shale oil and gas evaluation aims to identify a sweet spot location for shale development by merging geological and engineering features, yielding integrated geological engineering development.

The properties of shale, such as its total organic content, mineral composition, and other physical characteristics, play a crucial role in the successful design and execution of hydraulic fracturing operations. Proper evaluation of these properties helps determine the production potential of shale. However, limited literature is available on using artificial intelligence to predict mineralogy. This lack of research motivated us to conduct a study on Formation G Shale of the Ordos, as shown in Fig. 1.1.

This study uses readily available conventional logs and XRD data from core logs to anticipate Formation G total organic content and mineralogy, specifically clay, feldspar, quartz minerals, and porosity. Both minerals are vital constituents of shale and aid in determining the brittle and ductile zones within the shale formation.

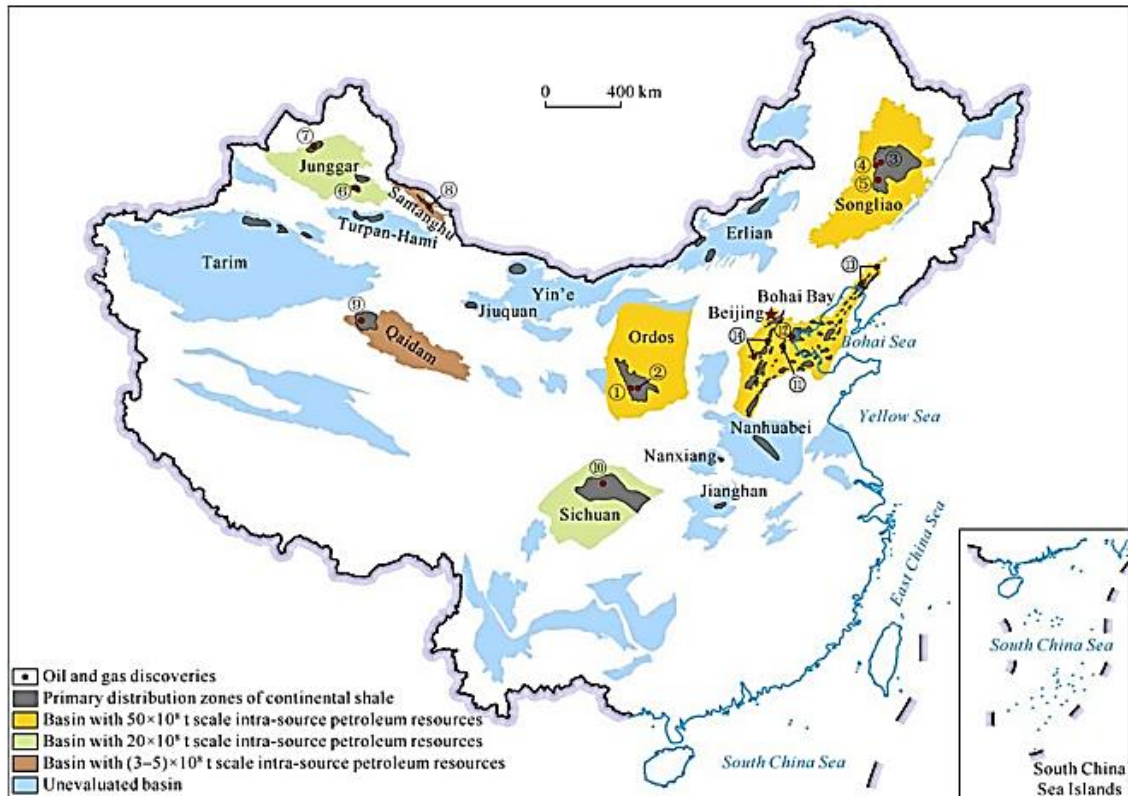


Fig. 1.1: Plane Distribution showing the Ordos Basin and distribution map of continental shale oil in four basins (modified after (Jiao, 2020))

1.2 Geological settings of the study area

The study area of this work is the Formation G shale reservoir of Ordos. The Ordos Basin, which includes the provinces of Shaanxi, Ningxia, Gansu, Inner Mongolia, and Shaanxi, is China's second-largest oil and gas-bearing basin. As a result of Indonesian movement and tectonics in the Late Triassic, the Yanchang formation in the basin formed a complete and typical clastic sedimentary system with fluvial lacustrine facies. Chronologically, the sedimentary cycle separates into ten oil-bearing formations (Formation J - Formation A) from bottom to top. The Formation J - Formation G period is primarily the lake's expansion period, with the shoreline rapidly moving outward.

The lake basin reached its peak during the Formation G_3 period. The primary source rocks of the Mesozoic basin were widely deposited deep lake mudstone. After a brief period of lake flooding, the basin entered a shrinking phase during the Formation F - Formation J period.

The lake basin gradually vanished from the Formation C - Formation A route. Several favorable source-reservoir-cap compositions resulted in the vertical creation of many oil-bearing series within the reservoir. Fig. 1.2.

The Jiyuan Oilfield is located between the secondary tectonic units Yishan Slope and Tianhuan Sag in the western Ordos Basin. It was in the sedimentary core of the lake basin during the Triassic Formation G period. The water depth is significant, and deep lake-semi-deep lake surfaces have developed. The basin's primary rock source is vast strata of lacustrine, organic-rich black mudstone (Fig. 1.2). According to core observations, Formation G primarily produces carbonaceous mudstone containing plant bits, fish scales, and fish stone. The hydrocarbon source rock is often thicker than 40 meters. Sapropel Formation and vitrinite predominate in the kerogen, particularly kinds II and I. Formation G shale minerals include clay minerals (such as illite, smectite, and chlorite), quartz, feldspar, and carbonate. These minerals contribute to the mechanical properties of the reservoir, pore structure, and brittleness of the shale reservoir.

Natural fractures are a prominent feature of the Formation G shale reserve. These fracture's orientation, distribution, and connectivity alter the formation's fluid flow and migration pathways, making them a key consideration in exploration, exploitation, and production strategies.

Finally, the geology of the Formation G formation in the Ordos Basin provides a vast collection of information about ancient ecosystems and hydrocarbon potential. Its organic-rich shale, mineralogical composition, natural cracks, and depositional history add to its importance as a shale oil and gas source. As energy exploration and extraction evolve, Formation G remains a hotspot for research and innovation in conventional hydrocarbon resources.

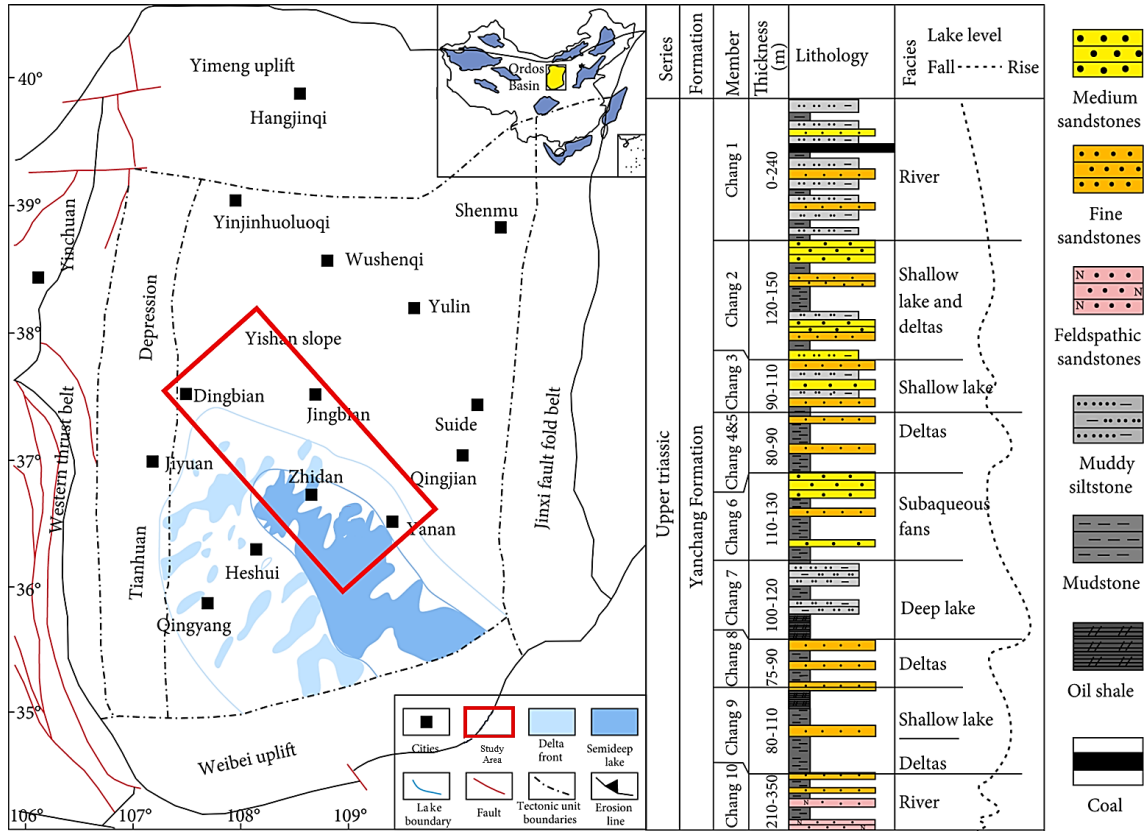


Fig. 1.2 Location and stratigraphy of the study area (modified after (Ma, 2021))

1.3 Problem statement

Formation G, a prominent part of the Mesozoic Yanchang Formation in the Ordos Basin, has attracted much attention because of its possible use as a reservoir for shale oil. The Formation G shale reservoir has excellent potential. However, its mineral composition still needs to be better understood and characterized, making it challenging to utilize and manage fully shale formations such as the Formation G shale. Precise assessment is critical because it directly affects the mechanical characteristics of the reservoir, its capacity to produce hydrocarbons and the efficiency of hydraulic fracturing.

Traditional analysis methods, predominantly based on X-ray diffraction (XRD) and scanning electron microscopy (SEM), are effective, often time-consuming, and costly. They require physical samples that might not always adequately present the entire bulk reservoir. Moreover, these methods may only partially capture the heterogeneity and complexity of shale formations, leading to less precise characterization and evaluation.

The development of machine learning and sophisticated statistical techniques presents a viable remedy. The use of these methods, in particular, machine learning and multiple linear regression, to precisely forecast and assess the mineral content of shale reservoirs, such as clay, feldspar, quartz, and porosity in the Formation G formation, is still in its

infancy. More thorough implementation of these techniques is needed for predictive analysis, especially when dealing with intricate and diverse geosciences and petroleum engineering formations.

Furthermore, it needs to be clarified how Formation G relates to geological characteristics, including thermal maturity, total organic content, and total porosity. Machine learning and multiple linear regression must be researched and used to assess the shale reservoir. This lack of understanding makes it extremely difficult to forecast how the reservoir will behave when different extraction methods, such as hydraulic fracturing, are utilized.

Therefore, there is a pressing need for:

1. Utilizing machine learning and multiple linear regression to build and validate a robust predictive model for accurately assessing the Formation G shale deposit.
2. Determining how geological factors influence Formation G's mineralogical characteristics and improving reservoir behavior predictability.
3. Evaluating the feasibility of these cutting-edge in terms of accuracy, efficiency, and cost-effectiveness in contrast to conventional analytical procedures.

Addressing these challenges is critical for optimizing the exploitation and management of the Formation G shale reservoir, with implications for enhancing energy production efficiency.

1.4 Research objectives

This research aims to assess and characterize shale reservoirs in Formation G comprehensively.

The research objectives of this research are;

1. To determine the Formation G shale reservoir's Total Organic Carbon (TOC) using the Passey Method, MLR Method, Machine Learning, and Stacking Model.
2. Using the Multiple Linear Regression (MLR) model to predict and evaluate mineral composition in the Formation G shale reservoir.
3. To develop Machine Learning (ML) models to improve the prediction and evaluation of mineral distribution in the Formation G formation.
4. Assessing the impact of geological features such as porosity on mineral composition in Formation G shale formation.
5. Performance and comparative analysis of the MLR and ML models in the mineral characterization of the Formation G shale reservoir formation.

To achieve the objectives of this work, conventional data and XRD data from core samples from Formation G formation of Ordos are used, considering the lithological variety of the targeted reservoir. Traditional data will be utilized to determine the TOC using the Passey method, multiple linear regression, and machine learning methods such as gradient boost, adaptive gradient boost, random forest, and stacking technique. In contrast, multiple linear regression and machine learning methods such as gradient boost and extreme boost will be used to study the composition of mineral content in the Formation G reservoir.

1.5 Research Significance

The significance of this research work is as follows;

1. Characterize the shale reservoir of Formation G by employing the Passey, MLR, and Machine Learning Methods for TOC estimation to critically understand the hydrocarbon source potential.
2. Estimate and characterize the clay, feldspar, quartz, and porosity of the shale reservoir using multiple linear regression and Machine Learning methods such as Gradient Boost and XGBoost to refine, characterize, and identify nuanced geological features that traditional methods might overlook.
3. This research is the vanguard for adopting analytical techniques in geological and geophysical studies for future exploration and development strategies in the oil and gas sector.
4. This study will help identify the best zone for hydraulic fracturing design, as shale's mineralogy and geologic features are the primary features for defining brittle zones/ layers for fracturing operations.
5. To develop a model to predict and evaluate mineral composition and porosity from conventional data.

1.6 General Methodology and Workflow

Core analysis is the most efficient means of assessing shale oil reservoirs, although the expense of coring and testing makes thorough well evaluations impossible. Despite this, every well possesses conventional logs. Therefore, the TOC, mineral composition, and porosity will be calculated and evaluated by integrating core and conventional logging data for the Formation G shale oil reservoir. Fig. 1.3 Shows the intended workflow for the methodology in my research.

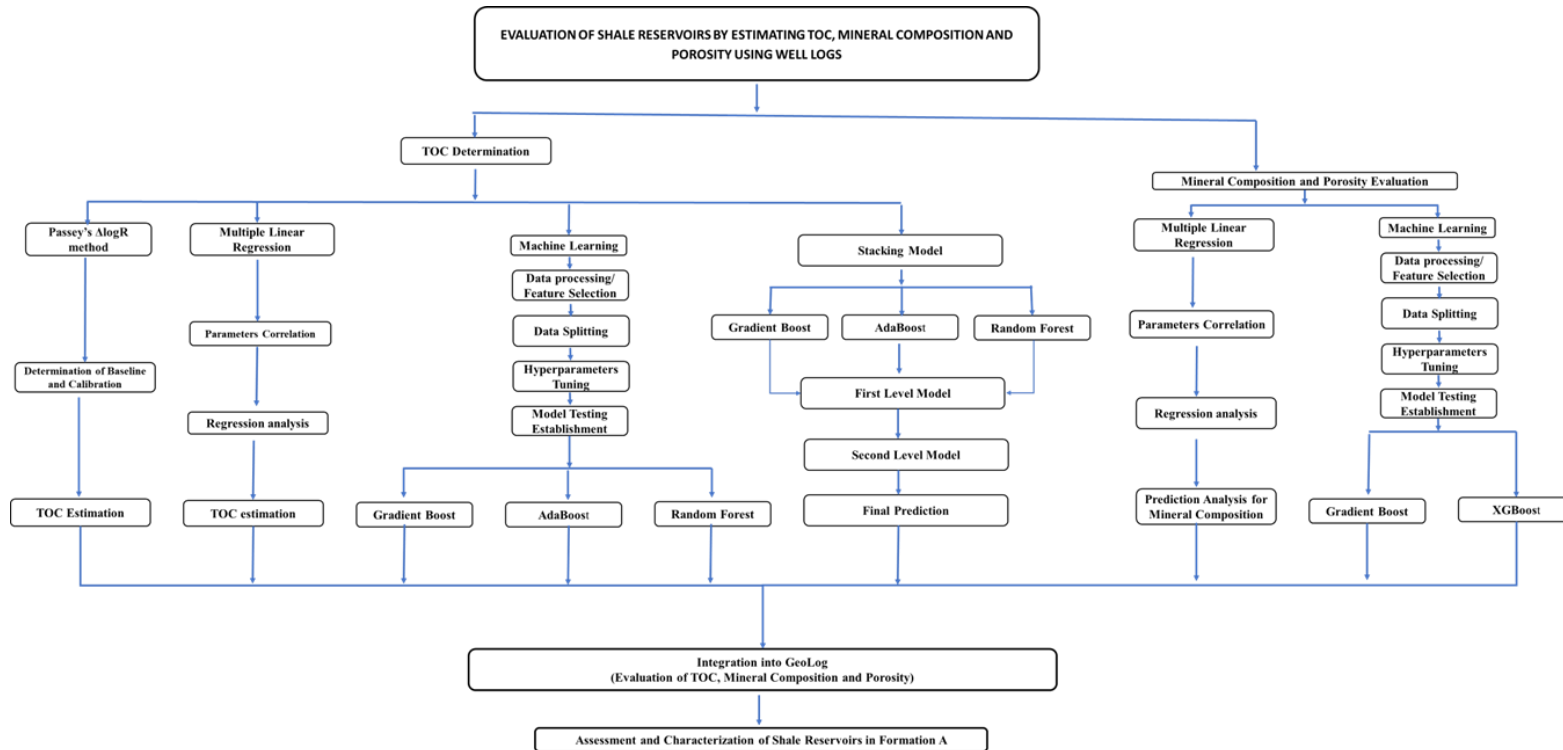


Fig. 1.3: Computational workflow developed for the research

Chapter 2 Literature Review

2.1 Shale Gas as an Energy Source: Challenges and Advances

Natural gas has become an efficient and increasingly favored energy source compared to petroleum and coal (Wanniarachchi, 2015), with an annual consumption growth rate of 0.8% (Conti J., 2014). Unconventional gas resources, such as tight gas sands, coal bed methane, and shale gas, offer vast potential (James W. Schmoker¹, 2002; A. A. Radwan & Nabawy, 2022; A. E. Radwan, 2022). The Middle East and North Africa rank third globally, with an estimated 2547 trillion cubic feet of shale gas resources (Kawata & Fujita, 2001). However, estimating shale gas reserves is fraught with uncertainties due to numerous factors requiring diverse methods for precision (A. A. Radwan & Nabawy, 2022). Adequate geological characterization of shale gas reservoirs includes analyzing lithofacies, sedimentology, geochemistry, porosity networks, and fractures (Josh, 2012). Shale volume estimation is a cornerstone in the oil and gas industry, serving as a key parameter in predicting petrophysical characteristics such as effective porosity and water saturation (Elmahdy, 2023; Gogoi & Chatterjee, 2019). Given the prohibitive costs and limited availability of core data, these predictions rely heavily on well-log data (Li & Han, 2017). Gamma-ray (GR) and spontaneous potential (SP) logs are the most used methods for differentiating between clean and shale layers, as shales typically exhibit higher radioactivity due to their elemental composition (Chen, 2017; Josh, 2012).

While conventional methods have been effective, they often fall short in accounting for reservoir heterogeneity. Relying on one or two parameters from log data introduces uncertainties and inaccuracies, especially in complex formations (Szabó, 2011). Machine learning (ML) approaches, which analyze multiple parameters and uncover intricate relationships within the data, offer a promising alternative for more reliable shale volume estimation (Amosu & Sun, 2021; Vikara, 2020).

2.2 Mineralogical and Mechanical Properties of Shale Gas Reservoirs

Shale is a fine-grained clastic sedimentary rock that consists of clay mineral composition and other minerals and particles in layer size, mainly quartz and calcite. These rocks are fine-granular and layered, with isolated porosity and low permeability (Nhabanga, 2021). Petroleum geology considers shales containing organic materials to have a high potential as source rocks. Because of their poor permeability, shales can also be used as cap rocks to prevent oil leakage in oil traps (Venieri, 2021).

Critical minerals in shale, such as quartz and clay, significantly influence production potential (Huang, 2013). Fracture intensity, organic content, porosity, adsorbed gas, and brittleness control reservoir productivity (Ouenes, 2014). Low permeability is a significant challenge, requiring advanced techniques like hydraulic fracturing to achieve economic flow rates (Daigle & Screaton, 2012; Kumar, 2016). Hydraulic fracturing involves injecting high-pressure fluids to create fracture

networks, enhancing gas flow by increasing pore space and reducing tortuosity (Boz, 2014; Ghahremani & Clapp, 2014). However, water-based fracturing fluids cause challenges, including long cleanup times, formation damage, and environmental concerns (Arnold, 1998; Nicot & Scanlon, 2012).

Clay minerals in shale formations exacerbate these issues due to their water absorption, leading to swelling, reduced pore space, and hindered gas movement (Rahman, 2007; Zhou, 1996). High clay content imparts ductility, reducing hydraulic fracturing effectiveness (Ding, 2012; Jarvie, 2007).

2.3 Shale Brittleness Index and Fracturing Design

The Shale Brittleness Index (BI) categorizes ductile and brittle zones in shale formations based on mineralogical composition (Jarvie, 2007). High quartz content indicates brittleness, while clay-rich zones are ductile (Safari, 2013). Mineralogical and geomechanical properties are critical for optimizing hydraulic fracturing operations (Jacobi, 2009). Zones with high brittleness and favorable geomechanical properties enhance gas recovery (Abouelresh & Slatt, 2012).

2.4 Challenges in Mineralogical Evaluation

Quantitative mineralogical determination in shales, such as clay, feldspar, and quartz content, is crucial for identifying high-potential zones for hydraulic fracturing. However, these evaluations are expensive, time-consuming, and often inconsistent across laboratories (Clarkson, 2013). Additionally, the absence of cores limits reservoir-wide mineralogical data availability.

2.5 Data-Driven Approaches in Shale Reservoir Evaluation

Advancements in artificial intelligence (AI) and machine learning (ML) offer solutions to these challenges. Data-driven methods process vast datasets to predict mineral composition, porosity, and other reservoir properties more efficiently than traditional techniques (Yu, 2020). The International Energy Agency (International Energy Agency (IEA), 2017) predicts a 5% increase in reserves and a 10-20% reduction in production costs through digital technologies.

Supervised ML models use well-log inputs, such as Gamma Ray (GR) and Density (DEN) logs, and outputs from X-ray diffraction (XRD) analyses to train models (Abudeif, 2016). Unsupervised learning identifies patterns in unlabeled datasets, enabling novel insights into shale reservoirs (Srivardhan & Pal, 2015). ML techniques also integrate direct (e.g., XRD, SEM) and indirect methods (e.g., Spectral Gamma Ray Logs) for comprehensive shale characterization (Elmahdy, 2023).

2.6 Applications of AI in Shale Gas Development

Data-driven approaches provide fast, accurate solutions for reservoir challenges. For example, ML algorithms identify sweet spots and predict production potential in shale formations (Chen, 2017;

Liu, 2019). These methods enable cost-effective evaluations without extensive laboratory data, aiding decision-making in unconventional reservoir development (Dong, 2018; Szabó & Dobróka, 2018).

2.7 Shale Gas in the Ordos Basin

The Ordos Basin, a prominent shale reservoir, has been the focus of extensive research to overcome its unique challenges. It is the second-largest sedimentary basin and a significant oil and gas producer, with annual hydrocarbon production exceeding 7000×10^4 tons (Fu, 2020). Shale oil from Formation G Member is primarily self-sourced, with oil stored in adsorbed and free states within source rock sequences (Fu, 2020). Characterizing these reservoirs is essential for efficient production.

Researchers have emphasized integrating well log and seismic data to improve reservoir characterization (Li & Han, 2017). Zhao et al. (2016) utilized spectral gamma-ray logs to estimate total organic carbon (TOC), while Song et al. (2021) explored micro-resistivity and dielectric dispersion logs for permeability estimation. These advanced techniques provide critical insights into shale porosity and fluid saturation as fracture characterization remains a crucial area of study.

Wang & Chen (2019) employ borehole image logs to analyze natural fractures, underscoring their role in enhancing permeability and optimizing hydraulic fracturing strategies. K. Liu (2021) introduced nuclear spectroscopy tools for mineral identification and porosity estimation, showcasing their ability to address the complexities of shale mineralogy.

Nuclear magnetic resonance (NMR) logs have also gained traction. Ma et al. (2021) analyzed T1/T2 relaxation times in the Ordos Basin, identifying fluid-filled pore spaces and improving porosity and permeability estimates. Zhang (2019) combined image logs and stress measurements to enhance fracture distribution and stress orientation analysis, providing valuable insights for hydraulic fracturing designs.

2.8 Shale Volume and Petrophysical Characteristics

Estimating shale volume is vital in the oil industry to predict petrophysical parameters like effective porosity and water saturation. Core data, when available, validates well-log relationships, though gamma-ray (GR) and spontaneous potential (SP) logs are commonly used due to their sensitivity to shale (Zou, 2013). However, traditional methods can be error-prone due to reservoir heterogeneity. Advanced machine learning (ML) approaches offer better accuracy by incorporating multiple log parameters.

2.9 Innovative Approaches in Shale Volume Estimation

Using a nonlinear exponential relationship, Szabó (2011) developed a statistical method that links gamma-ray data to shale volume. This method provides a general formula for defining a linear

relationship, making it an independent reference for shale volume estimation. Building upon this framework, (Szabó & Dobróka, 2018) employed neural networks that integrated density, gamma-ray, and P-wave velocity logs, resulting in greater accuracy than traditional factor analysis methods. Furthermore, Szabó & Dobróka (2018) advanced the estimation of shale volume by utilizing genetic algorithms alongside Larionov's method. Additionally, Abordan and Szabó (2018) explored the application of Particle Swarm Optimization to refine factor analysis in the context of Hungarian reservoirs.

2.10 Machine Learning in Shale Reservoir Evaluation

AI-driven methods, such as Random Forest (RF), Gradient Boosting (GB), and Neural Networks, have shown promise in reducing uncertainties associated with conventional techniques. Studies highlight their effectiveness in integrating seismic and log data for comprehensive reservoir evaluation (Mustafa et al., 2022).

2.11 Applications in the Ordos Basin

Well logs and seismic data are crucial in accurately characterizing reservoirs, facilitating precise porosity estimation and fracture analysis. Numerous studies underscore the importance of integrating advanced log interpretations with machine learning techniques to effectively address the complexities associated with mineralogy and fluid interactions within these reservoirs.

2.12 Future Directions and Research Opportunities

Integrating lithology, mineralogy, and geomechanical properties is crucial for optimizing hydraulic fracturing. Enhanced computational methods and interdisciplinary approaches are needed to address heterogeneity, flow mechanisms, and fracture networks in unconventional reservoirs (Gale, 2014; Javadpour, 2007). Expanding AI applications in reservoir evaluation and exploring innovative ML techniques can further enhance shale gas development, ensuring sustainable energy production. Understanding shale properties, utilizing data-driven technologies, and optimizing hydraulic fracturing designs are crucial for unlocking the full potential of shale gas resources. By addressing challenges in mineralogical evaluation and reservoir characterization, we can advance unconventional hydrocarbon production, effectively and sustainably meeting the increasing energy demands.

2.13 Proposed Study: Formation G of the Ordos Basin

After reviewing various literature, research gaps in this field were identified for evaluating the mineral composition and porosity in the Formation G of the Ordos Basin. This work aims to bridge the gap between geological and petrophysical characterization of unconventional resources by refining methodologies and integrating data and technologies to unlock the full potential of shale reservoirs. Conventional and XRD data from core logs from Formation G of the Ordos Basin will

be used. This work will also guide the exploration and development of unconventional continental oil and gas resources, as the main objective of this work is to apply the machine learning-based prediction models to determine the content of three essential minerals (clay, feldspar, and quartz), porosity and TOC using the Passey's $\Delta \log R$ method, Machine Learning methods and Stacking Model in Formation G shale formation using continuous core and conventional logs suites of three wells to build a prediction model and also to compare the different ML methods and Multiple Linear Regression Method in evaluating the TOC, mineral composition and porosity. The viable statistical technique (Multiple Linear Regression) and Artificial Intelligence approaches (Gradient Boost (GB), Adaptive Gradient Boost, Random Forest, and Extreme Gradient Boost (XGBoost)) would be vital to obtaining quantitative information about the above-mentioned minerals and TOC. A prediction model will be developed using machine learning and MLR methods to estimate the new well mineral composition and TOC accurately. These values obtained from the new well will then be integrated into the petrophysical software to visualize the wireline data and increase the utility of these predicted insights. This integration allows for the display and interpretation of geologic formations through wireline logs, which are crucial for understanding the geophysical properties.

By feeding the petrophysical software with the predicted values from a newly developed model, one can easily visualize the mineralogical composition and TOC content, leading to more informed decisions regarding exploration and production strategies. This methodology has thus provided a significant advancement in digital pyrotechnical applications by integrating predictive models with standard geological software tools. The seamless integration of predictive modeling outputs with visualization tools such as petrophysical software indicates a rising trend in geophysics and geosciences digitization, increasing accuracy in identifying productive zones, monitoring reservoir quality, and optimizing field development plans. On the other hand, intelligent models provide a cost-effective and convenient method for determining the mineralogical composition of shale reservoirs.

The current study proposes using Passey's method, MLR, Machine Learning methods, and Stacking Model to evaluate TOC, as well as artificial intelligence and MLR models for forecasting the mineralogical composition of Formation G Shale, specifically the clay, feldspar, quartz content, and porosity.

Chapter 3 Evaluation of Total Organic Carbon

3.1 Introduction

Total Organic Carbon (TOC) is defined as total organic carbon in weight percent, including all kerogen and fluid hydrocarbons available in the matrix. It is one of the key reservoir quality indicators in shale reservoirs. This study will employ the Passey $\Delta\log R$, multiple linear regression, and machine learning methods to estimate the TOC in Wells 1. The TOC measurement in this work was obtained from core samples and complimented by XRD.

3.2 Methodology

3.2.1 The Passey Method

Passey initially proposed the $\Delta\log R$ method (Q. R. Passey, 1990) which has been widely used for TOC estimation using well logs in carbonates and clastic rocks. $\Delta\log R$, i.e., the sonic resistivity overlay plot AC and true resistivity R_t logs in one track; AC and R_t are scaled as a 50 $\mu\text{s}/\text{ft}$ ratio to one resistivity cycle. The separation between two curves (AC to the left and RT to the right) is defined as $\Delta\log R$ (Q. R. Passey, 1990). The AC curve will reflect low-density/low-velocity kerogens, while the R_t curve will respond to Formation fluid (Tenaglia, 2020). The method of sonic-resistivity overlay ($\Delta\log R$) is used to calculate TOC as follows:

$$\Delta \log R = \log \frac{R}{R_{\text{baseline}}} + K(\Delta t - \Delta t_{\text{baseline}}) \quad (3.1)$$

$$K = \frac{\log (R_{\text{max}}/R_{\text{min}})}{\Delta t_{\text{max}} - \Delta t_{\text{mn}}} \quad (3.2)$$

$$\text{TOC} = (\Delta \log R) \times 10^{2.297 - 0.1688 \text{LOM}} \quad (3.3)$$

where R is deep resistivity log (ohmm), Δt (AC) is the sonic transit time ($\mu\text{s}/\text{ft}$ or $\mu\text{s}/\text{m}$), R_{Baseline} and $\Delta t_{\text{Baseline}}$ are the resistivity and sonic transit time values at the baseline separately; LOM (Local Level of Organic Metamorphism) is a constant related to thermal maturity. Sometimes, the sonic transit time curve may not be unavailable, so density or neutron log can be used instead (Huang, 2013; Q. R. Passey, 1990). The scaling of sonic and resistivity well logs during overlaying is set as a ratio of $-164 \mu\text{s}/\text{m}$ ($-50 \mu\text{s}/\text{ft}$) sonic transit time (Δt) to one logarithm resistivity cycle $\Omega \text{ m}^{-1}$ (for instance, 1–10 ohmm) (Q. R. Passey, 1990).

TOC can be calculated using $\Delta\log R$ method, providing that the baseline of AC and RT is determined. Source rock intervals can be qualitatively identified using GR, CN, DEN, and resistivity logs.

Additionally, plotting AC and RT in one track with scaling of 100 $\mu\text{s}/\text{ft}$ (for instance, 140–40 $\mu\text{s}/\text{ft}$) transit time to two logarithm resistivity cycles (10 - 1000 ohmm) also gives the source rock intervals. As is known, AC and RT logs will overlap with each other in non-source rock intervals, while the two logs will deviate evidently at the source rock intervals. There are many $\Delta\log R$ based methods widely used to estimate TOC using well logs (Chen, 2017).

The limitation of $\Delta\log R$ is that the baselines vary in different wells (Ouenes, 2014). Additionally, the presence of pyrite will mask the resistivity profile and show low resistivity (Q. R. Passey, 1990).

3.2.2 Multiple Linear Regression (MLR) Method

Multiple Linear Regression (MLR) is a statistical method that models the relationship between multiple independent variables and a single dependent variable. By fitting a linear equation to observed data and analyzing the correlation with XRD data and porosity, we can select variables having low collinearity, significant correlations, and P-values less than 0.05. We refine our model by performing residual analysis by excluding variables with high collinearity or P-values above this threshold. Our prediction model for TOC, mineral composition, and petrophysical features in two wells is built using the equation obtained from MLR and integrated into petrophysical software, enabling enhanced geological and petrophysical interpretation. MLR is essentially an extension of standard least-squares regression, involving more than one explanatory parameter.

$$y = \beta_0 + \beta_1 x_1 + \dots \beta_i x_i \dots \beta_n x_n + \varepsilon \quad (3.4)$$

where y is the dependent variable; $x_1 \dots x_n$ are independent variables; β_0 is the y-intercept; β_i is the regression coefficient of the i th independent variables; ε is the modeling error, also known as the residual.

3.2.3 Machine Learning (ML) Method

In the rapidly evolving area of reservoir characterization, machine learning has emerged as a game-changing tool that provides sophisticated insights that traditional methods often disregard (Mustafa, 2022). Shale formations, with their inherent intricacy and variability, pose a significant challenge for precise mineralogical characterization, which is crucial in determining the quality and potential for reservoir extraction. The use of machine learning in this context is not simply innovative; it is essential for deciphering the intricate connections and non-linear relationships between geophysical log data and mineral components.

In this study, three (3) widely accepted machine learning techniques, such as the Gradient Boost (GBoost), Adaptive gradient boost (AdaBoost), Random Forest (RF), and the Stacking technique, are applied. Using machine learning techniques, GR, AC, NPHI, DEN, RD logs for about 335 feet in Well 1 thick shale intervals, carrying 102 data points were used to develop the machine learning predictive model for TOC in Formation G Shale formation. However, the Stacking technique takes the outputs of sub-models Gradient Boost (GBoost), Adaptive gradient boost (AdaBoost), and Random Forest (RF) as an input to predict the TOC and attempts to learn how to combine the input predictions best to make a better output prediction were used to predict the TOC of the organic-rich Formation G shale at Ordos Basin.

3.2.4 Gradient Boost Method

Gradient boosting is an ensemble-supervised machine learning algorithm used for classification and regression tasks. As discussed in the bagging/boosting comparison section, gradient boosting trains multiple models sequentially, assigning higher weights to instances with incorrect predictions.

Therefore, challenging cases are the focus during training. Sequential model training using gradient boosting will gradually minimize a loss function. The loss function is minimized similarly to an artificial neural network model with optimized weights. The predictions are compared with actual values after building the training model in gradient boosting. The difference between prediction and actual values represents the model's error rate. The model's error rate can now be used to calculate the gradient, essentially the partial derivative of the loss function. The gradient is used to find the direction in which the model parameters would have to change to reduce the error in the next round of training. Unlike a neural network model, where the objective function is to minimize a loss function in a single model, multiple models' predictions are combined in gradient boosting.

Therefore, gradient boosting uses some of the hyperparameters used in random forest/ extra trees and other hyperparameters, such as learning rate, loss function, etc., that are used in an ANN model.

3.2.5 Gradient Boost Prediction Model Setup

The simple flow chart designed for building a prediction model for gradient boost and the stages on how Gradient Boost was implemented using sci-kit-learn is presented in Fig. 1.3. In this stage, the goal is to use gradient boosting to train an ML model to evaluate TOC using input features like AC, GR, RILD, DEN, and NPHI logs from different wells, which will be merged to build a prediction model.

First, one launches Python's Jupyter Notebook for programming. Next, import all the necessary libraries and datasets into your Jupyter Notebook. Then, preprocess the imported dataset. After that, visualize the distribution, box, and scatter plots for the imported features. Check for outliers in the plotted distribution and box plots. If there are any, process the data until all outliers are eliminated. Once done, visualize the Pearson correlation coefficient.

Define the x and y variables and import the "train_test_split" by dividing the data into a 70/30 split for training and testing. Then, begin training the dataset. Import the "GradientBoostingRegressor" library and define the hyperparameters within this algorithm. "n_estimators" is defined as the number of boosting stages or, in other words, the number of trees in the forest. After defining the hyperparameters, proceed to the fit to "(x_train, y_train)" process, and predict the test data.

Obtain the coefficient of determination (R^2) for the training and testing data and adjust accordingly until you obtain a satisfactory R^2 . Visualize the training and testing actual versus the predicted

values. Obtain MAE, MSE, and RMSE for 70% training and 30% testing data. Finally, obtain the feature importance of each input variable and permutation feature importance.

The machine learning model has been constructed and is now prepared to accept the same features that were used during the creation of the prediction model from different wells as input. Next, the data from the new well we intend to predict will be imported into the model. The ML model will then be executed on this new data, resulting in a new predicted output.

Finally, the newly predicted output will be depth-matched and integrated into the petrophysical software to analyze and interpret the mineral composition and porosity further. The new predicted values will then be visualized and interpreted as a wireline log for better interpretation and comparison of the unseen portion of the reservoir.

3.2.6 Adaptive Gradient Boost Method

Adaptive gradient boosting or AdaBoost is another form of supervised ML boosting algorithm. In adaptive boosting, the weights on the sample observations that are challenging to classify or predict are increased by altering the distribution of the training data set. Unlike AdaBoost, where the sample distribution gets modified, in gradient boosting, the distribution does not get modified, and the weak learners train on the remaining errors of the strong learners. Remember that a weak learner is defined as a classifier with a strong correlation with true classification. In other words, it can correctly classify or predict slightly better than random guessing. On the other hand, a strong learner is a classifier with a stronger correlation with true classification.

3.2.7 Adaptive Gradient Boost Prediction Model Setup

Applying the same procedure to build the prediction model in Gradient Boost, you Train the dataset by importing the "AdaBoostRegressor" library, define hyperparameters, and proceed with the fit to the "(x_train, y_train)" process to predict the test data.

3.2.8 Random Forest (RF) Method

The random forest algorithm is an ensemble technique that combines multiple decision trees. A random forest usually has a better generalization performance than an individual decision tree due to randomness, which helps to decrease the model's variance. Other advantages of random forests are that they are less sensitive to outliers in the dataset and don't require much parameter tuning. The only parameter in random forests that we typically need to experiment with is the number of trees in the ensemble. The basic random forest algorithm for regression is almost identical to the random forest algorithm for classification. The only difference is that we use the MSE criterion to grow the individual decision trees, and the predicted target variable is calculated as the average prediction across all decision trees.

3.2.9 Random Forest (RF) Prediction Model Setup

Applying the same procedure to build the prediction model in Gradient Boost, you Train the dataset by importing the " RandomForestRegressor" library, define hyperparameters, and proceed with the fit to the "(x_train, y_train)" process to predict the test data.

3.2.10 Stacking Technique

This ensemble technique works by applying input of combined multiple weak learners' predictions and Meta learners to achieve a better output prediction model. In stacking, an algorithm takes sub-model outputs as input and attempts to learn how to combine the input predictions best to make a better output prediction.

Stacking, also known as a stacked generalization, is an extended form of the Model Averaging Ensemble technique in which all sub-models equally participate according to their performance weights and build a new model with better predictions. This new model is stacked on top of the others, which is why it is named Stacking.

3.2.11 Architecture of Stacking

The architecture of the stacking model is designed to consist of two or more base/learner models and a meta-model that combines the predictions of the base models. These base models are called level 0 models, and the meta-model is known as the level 1 model. So, the Stacking ensemble method includes original (training) data, primary level models, primary level prediction, secondary level model, and final prediction. The basic architecture of stacking can be represented as shown below the image.

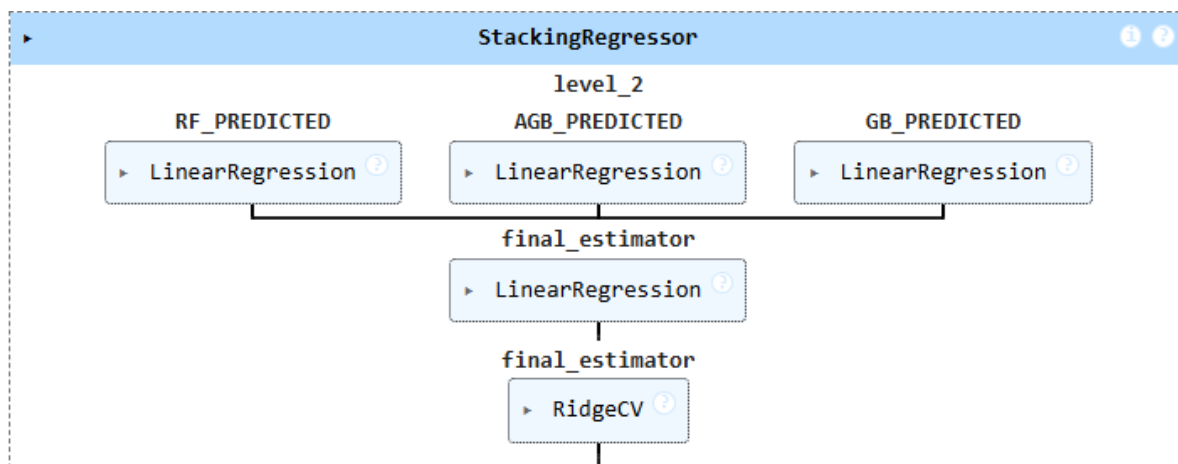


Fig. 3.1: The architecture of stacking shows the various levels of stacking.

3.3 Results of TOC evaluation in Well 1

The TOC in Well 1 was determined separately using Passey log R (Q. R. PASSEY, 1990), MLR, Machine Learning methods, and Stacking technique.

Using the Passey Method, the TOC was established with the following equation.

$$TOC = (\Delta \log R) \times 10^{2.297-0.1688LOM} \quad (3.5)$$

An equation to calculate the TOC using the MLR method was developed, and a model was established. The established model will then be integrated into the logging software to gain a wider view of the wireline interpretation and carefully observe the TOC in Well 1. The established model is as follows.

$$TOC = -0.367 + 0.002AC + 0.067DEN + 0.001CN - 2.30GR + 0.001RS \quad (3.6)$$

Fig. 3.2 presents the interpretation of TOC logging for Well 1 using Passey, MLR, ML, and the Stacking methods. The shale formation, developed between depths of 2054m - 2154m, was analyzed using GR, CAL, and SP logs. High GR readings are observed at depths of 2070m - 2096m, 2114m - 2142m, and 2147m - 2154m, indicating the presence of shaly rock. The calculated TOC contents show variability that may correlate with changes in lithology, as interpreted from the GR log, with high values observed in intervals from 2053m - 2055m, 2060m - 2095m, 2114m - 2142m, and 2147m - 2154m. The shale reservoir exhibited "five high and two low" characteristics on the conventional logging response curve, with high values of GR, CN, AC, RD, and RS and low values of SP and DEN at high depths of TOC. According to the logging response characteristics of shale oil reservoirs, conventional logging curves can be used to identify shale reservoirs, as shown in Table 3.1.

Table 3.1 Response parameters of shale oil from conventional wireline logging data in Well 1.

	AC (μs/ft)	CN (%)	DEN (g/cm ³)	GR (API)	RD (ohmm)
Minimum	55.58	8.04	1.99	46.29	12.68
Maximum	115.95	46.41	2.79	452.84	1008.93
Mean	73.31	20.34	2.63	154.38	97.45

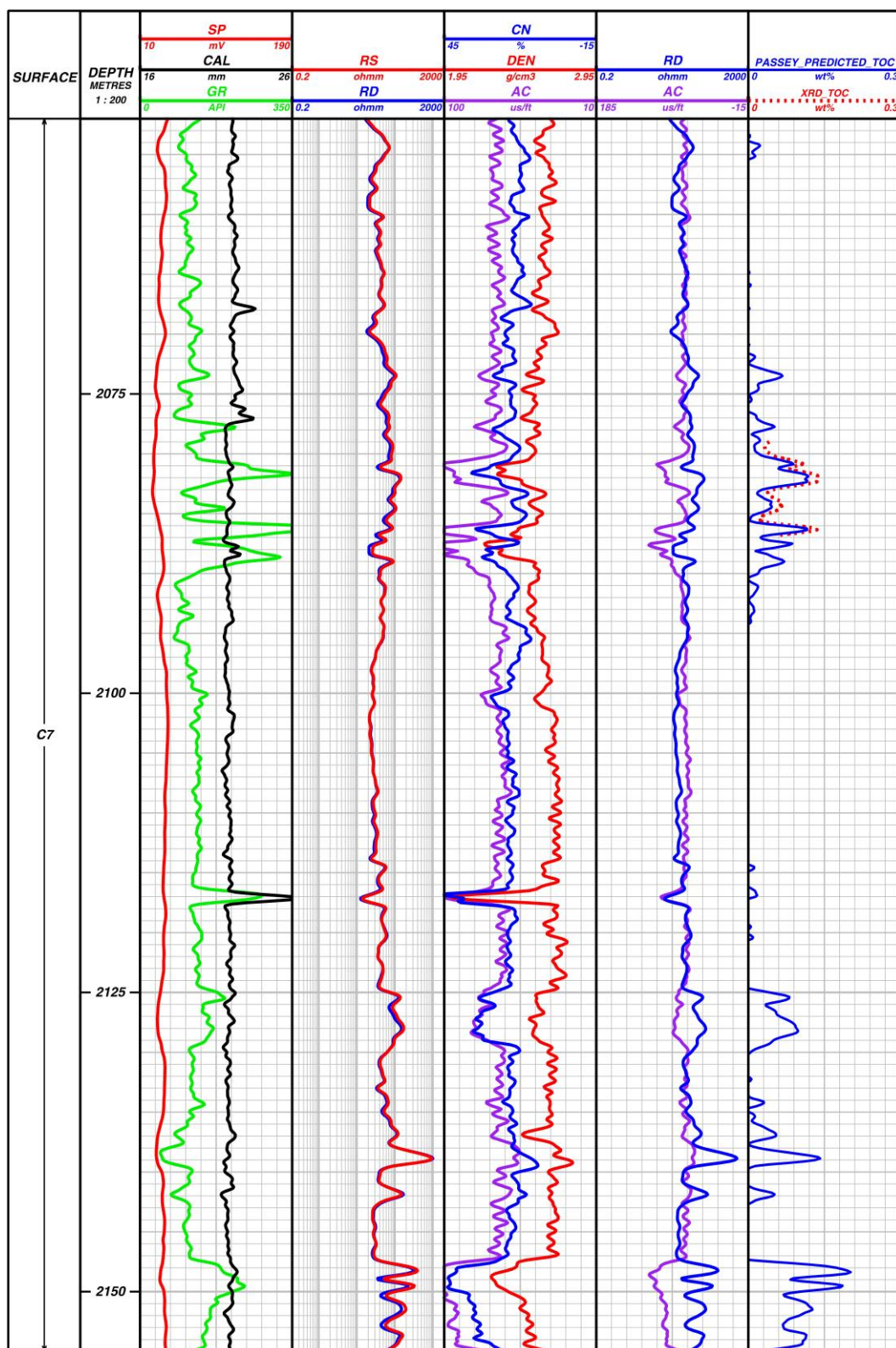


Fig. 3.2: Well log interpretation of lithology log, resistivity log, porosity log, and TOC derived from Passey's method in Well 1.

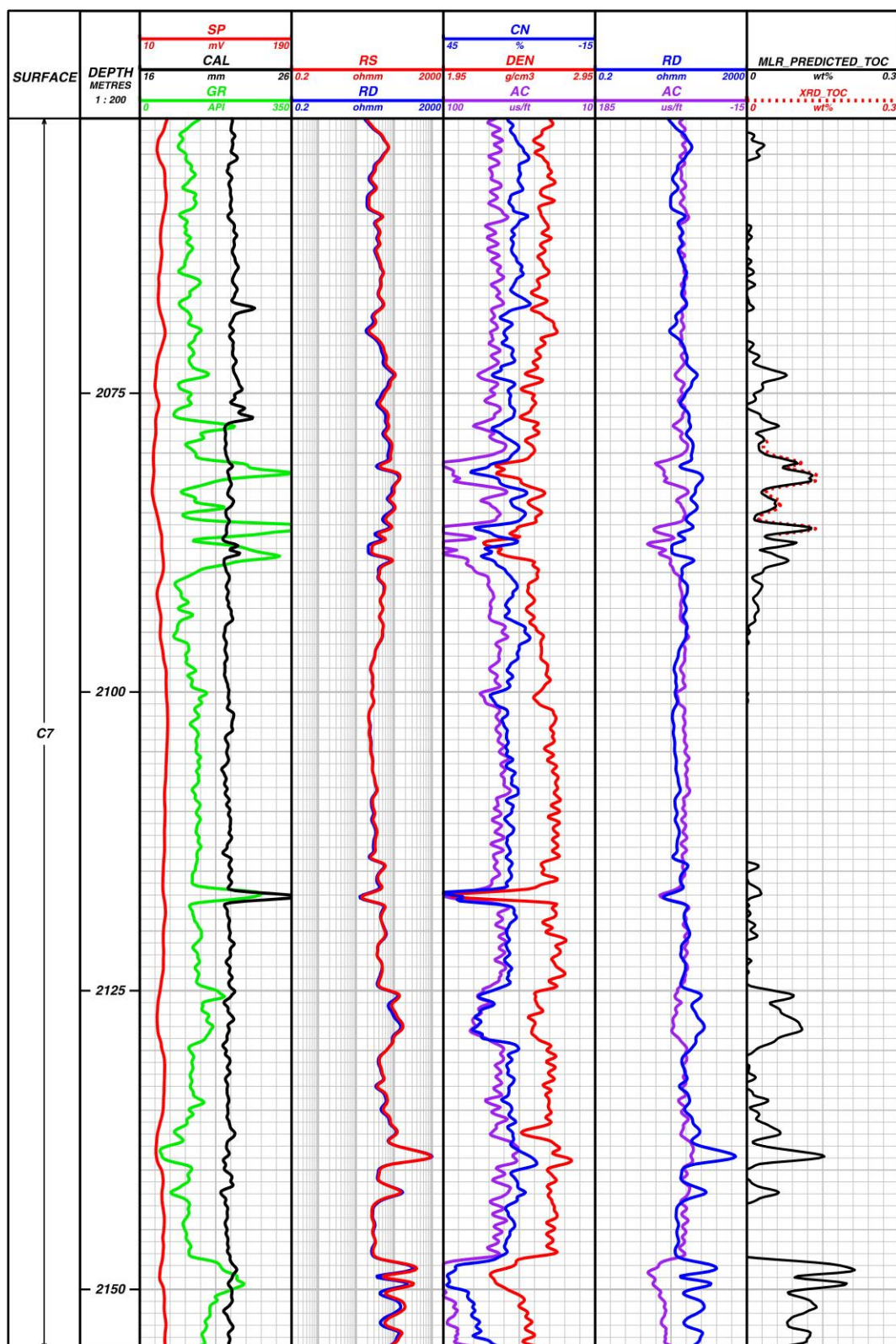


Fig. 3.3: Well log interpretation of lithology log, resistivity log, porosity log, and TOC derived from the MLR method in Well 1.

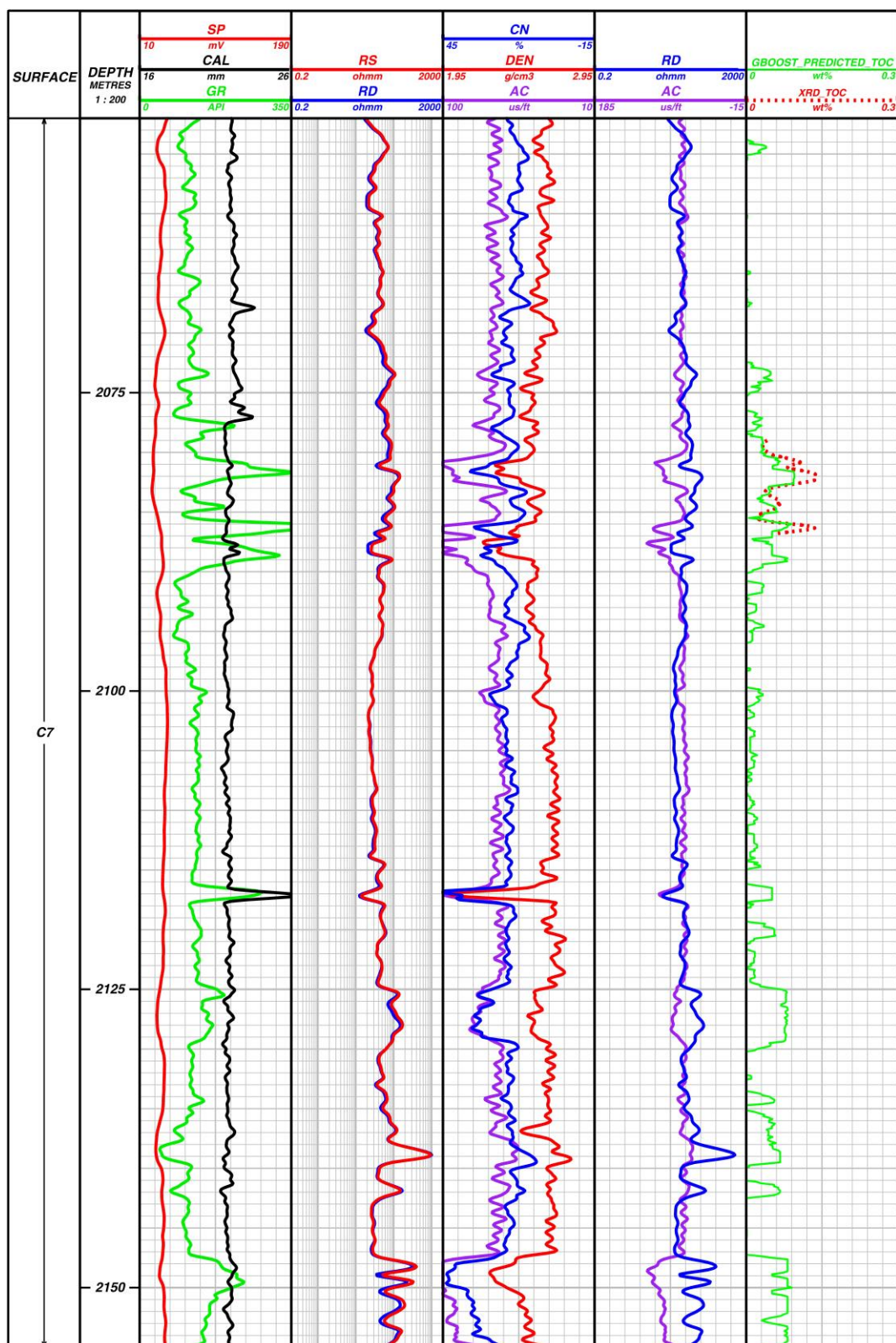


Fig. 3.4: Well log interpretation of lithology log, resistivity log, porosity log, and TOC derived from the Gradient Boost method in Well 1.

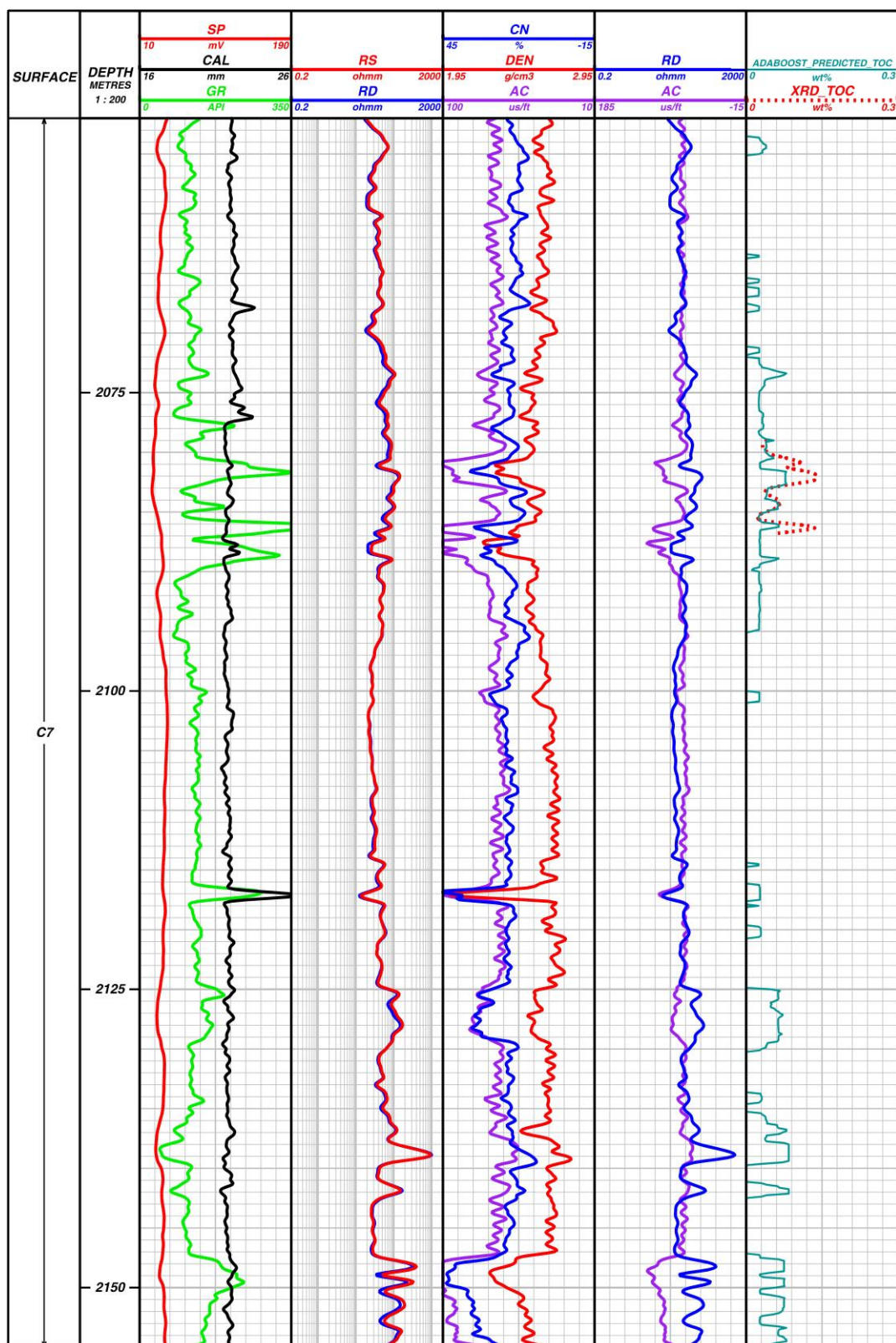


Fig. 3.5: Well log interpretation of lithology log, resistivity log, porosity log, and TOC derived from the Adaptive Gradient Boost method in Well 1.

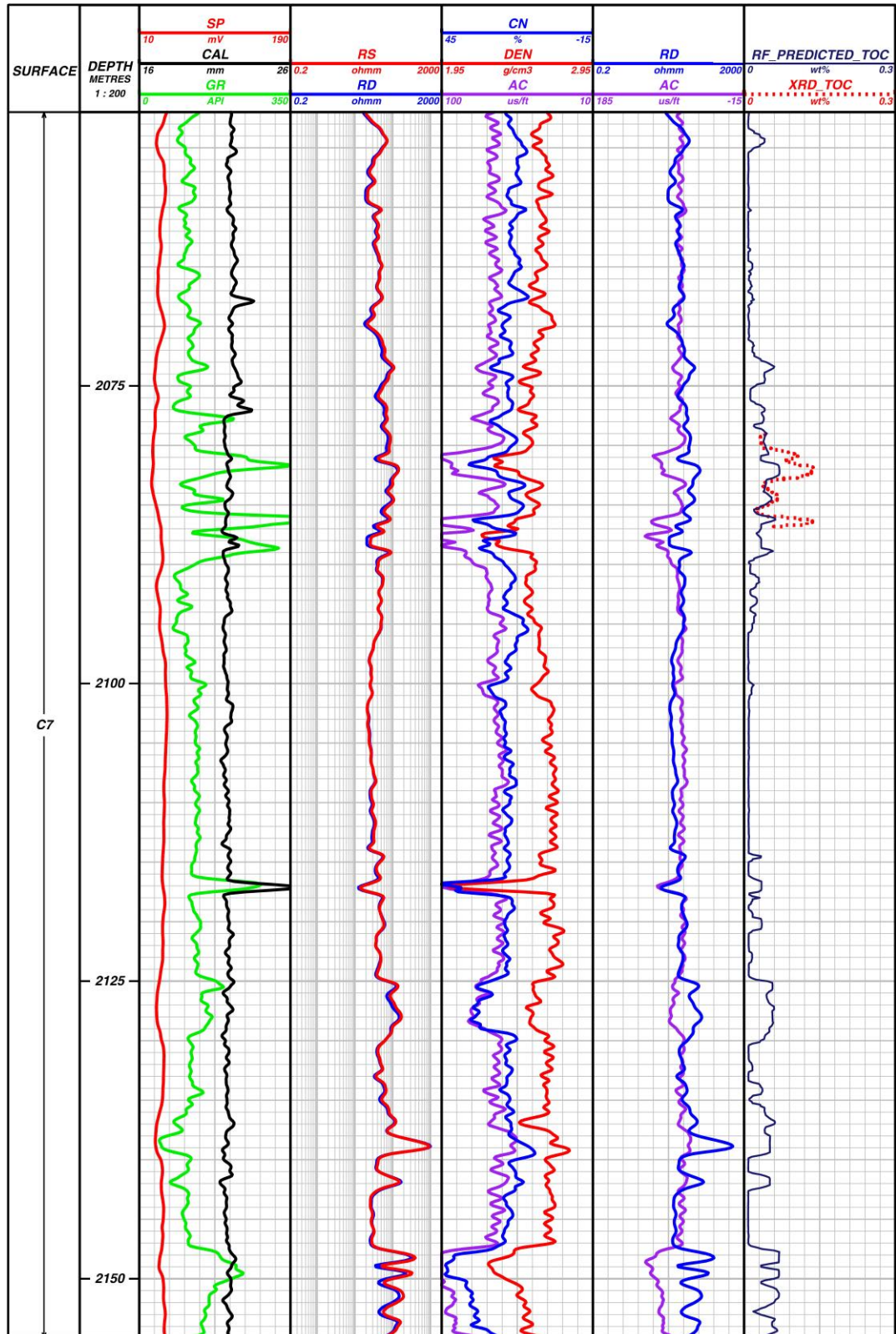


Fig. 3.6: Well log interpretation of lithology log, resistivity log, porosity log, and TOC derived from the Random Forest method in Well 1.

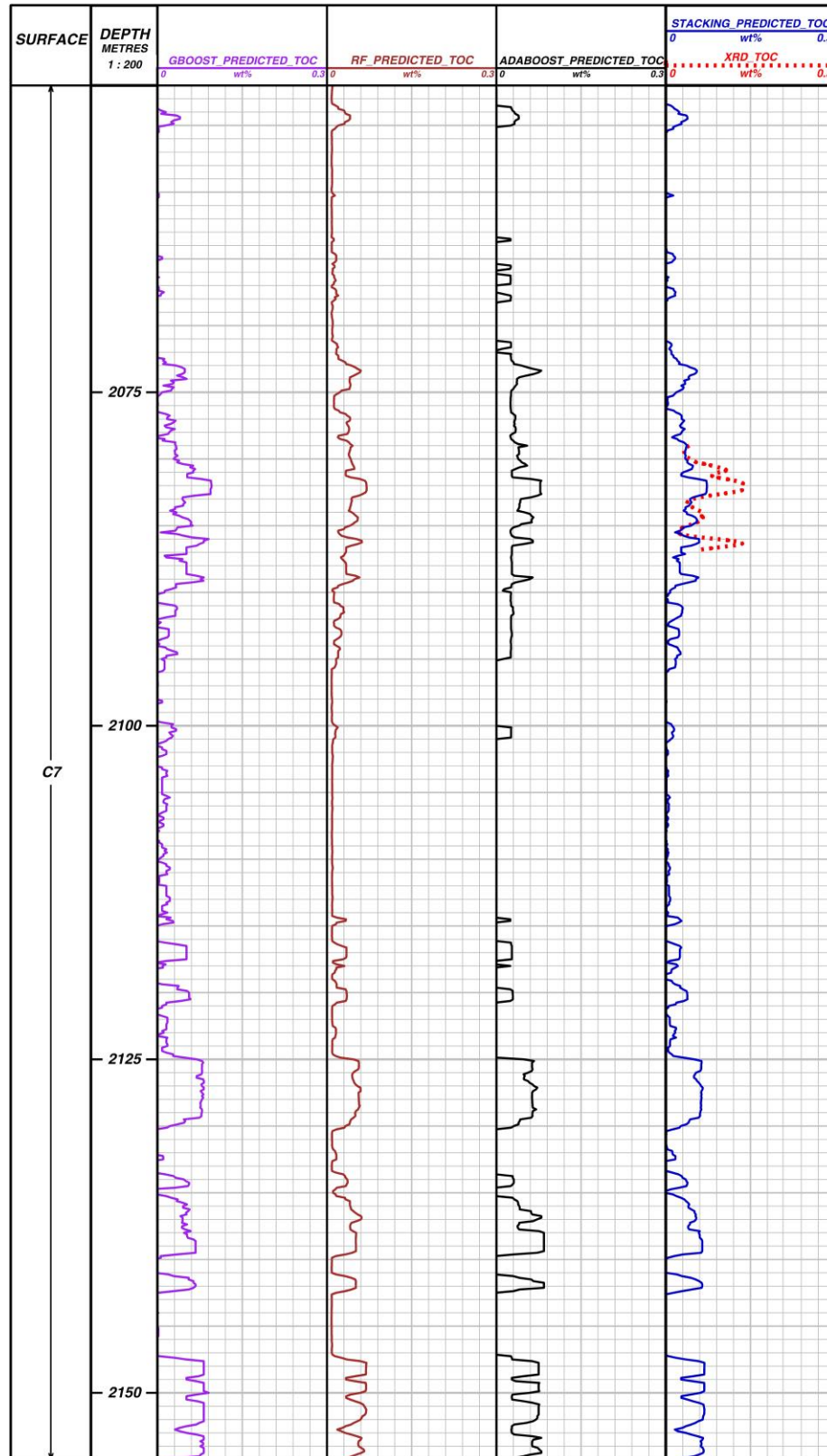


Fig. 3.7: Well log interpretation of TOC derived from the combination Gradient Boost, AdaBoost and Random Forest Prediction from Well 1.

3.4 Performance and Comparative Analysis of All Methods for Well 1

In Well 1, TOC was analyzed using various methods, and the results are depicted in Fig. 3.8. The predicted curves for the TOC using Passey's, MLR, Gradient Boost, Adaptive Gradient Boost, Random Forest, and the Stacking methods were in good agreement with the TOC from the rock eval pyrolysis. The predicted curve for the TOC obtained from the Passey and MLR methods was accurately predicted as they were in very close agreement with the actual XRD TOC from core sample than the machine learning methods.

However, the predicted output from the machine learning methods (Gradient Boost, Adaptive Gradient Boost, Random Forest) was used to build a model by the process of stacking the output obtained from them to train sub-models as input and attempts to learn how to combine the input predictions best to make a better output prediction. Thus, it can be observed that the stacking method presented output prediction that agrees with the actual TOC from XRD within some zones and continued all through the depth. The stacking of the various machine learning methods tends to give us a clearer picture of the performance of the various machine learning methods as the weak meta learners, in which all sub-models equally participate as per their performance weights and build a new model with better predictions. Overall, the auspicious results provide valuable insights into the TOC in Well 1. Fig. 3.8 displays the overall performance and comparison of all the methods and models.

3.5 Chapter Conclusion

In Well 1, Passey's method, MLR, Gradient Boost, Adaptive Gradient Boost, Random Forest, and Stacking methods were applied in Well 1 to evaluate the TOC. The Passey and MLR methods were more in close agreement with the actual TOC than the machine learning and the stacking technique. Meanwhile, when all the prediction curves from the entire methods approached the depth where the TOC core was measured, the predicted curve for the Passey method and MLR method showed good agreement with measured TOC, while the machine learning and the stacking method exhibited a relatively close agreement as seen in Fig. 3.8.

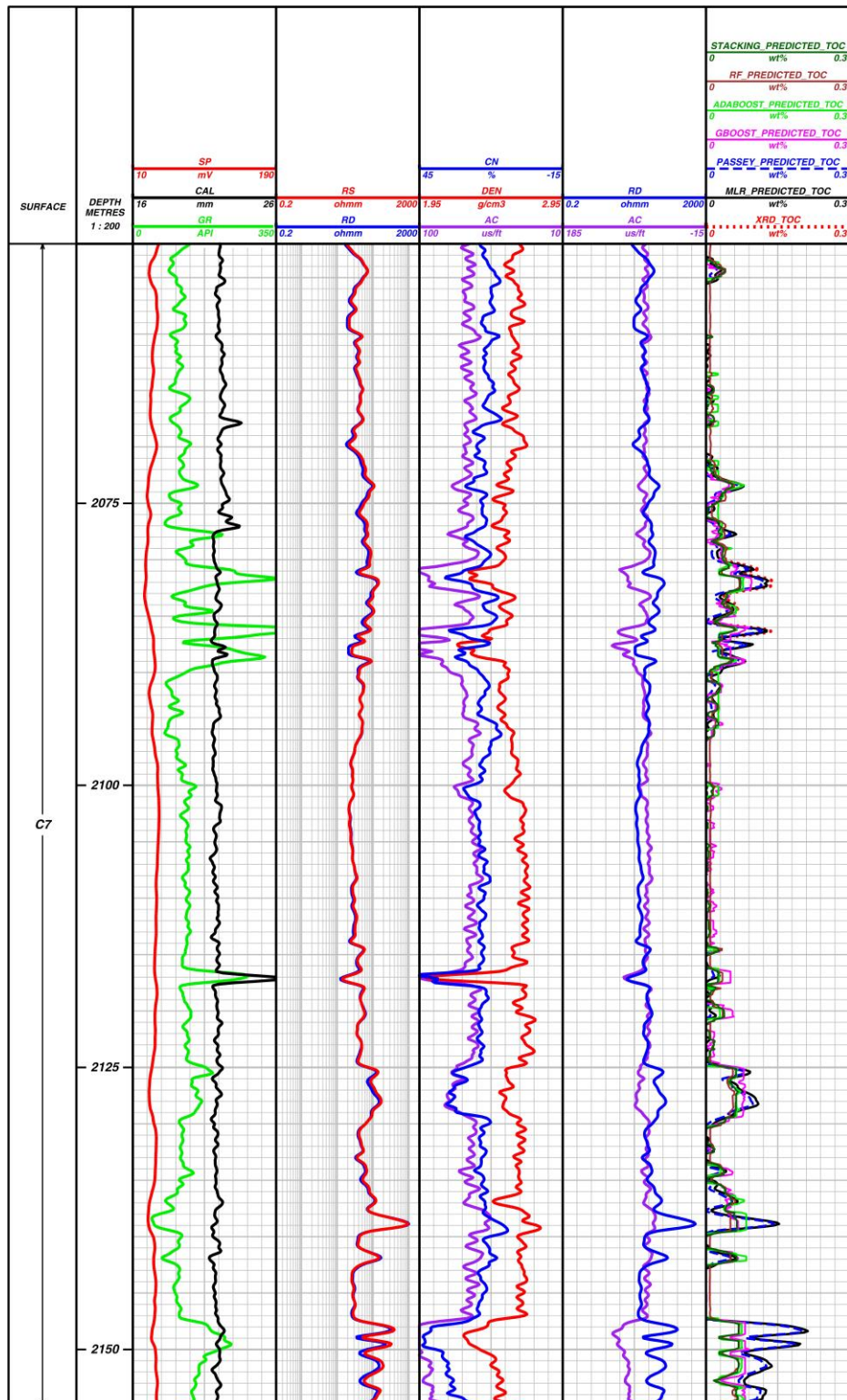


Fig. 3.8: Well log interpretation of lithology log, resistivity log, porosity log, and TOC derived from the Passey, MLR, ML and Stacking methods in Well 1.

Chapter 4 Evaluation of Mineral Composition and Porosity using Multiple Linear Regression (MLR) Method.

4.1 Introduction

Understanding mineral composition in shale reservoirs is crucial for several reasons of great importance for both the exploration and production stages of unconventional resources. Calculating mineral composition is a complex and essential work in evaluating shale reservoirs as it presents comprehensive information and knowledge about shale reservoir mineralogy. In evaluating the mineral composition in this study, a multiple linear regression method and artificial intelligence approach are employed to predict clay, feldspar, and quartz, including the porosity, using conventional logs of the wells penetrated at the Formation G shale in the Ordos Basin.

4.2 Methodology

The methodology adopted for this work can be seen in Chapter 3.2.2 and the computational workflow describing the stages of the processes involved in the statistical method of evaluating the mineral composition and porosity can be seen in Fig. 1.3.

4.3 Results

4.3.1 Mineral Composition Results from MLR Method in Well 1

In Well 1, reservoir features from well logging data such as the AC, GR, DEN, RD, RS, CN, and PE logs were utilized as the input variables. At the same time, clay, feldspar, and quartz minerals, including porosity, were the output variables. The correlation coefficient matrix for different log parameters is shown in Supplementary Appendix I (Table 4.1). Input parameters exhibited a good correlation with clay, feldspar, and quartz contents, including porosity, and were thus used to develop a prediction model. Input parameters have positive correlation coefficients with clay minerals content except for GR, RS, and PE log, which has a negative correlation with the clay content; the feldspar has a positive correlation coefficient with most input parameters except for GR and RD, which has a negative correlation coefficient with feldspar. The quartz content positively correlated with most input parameters except for RS, CN, and PE logs, which exhibited a negative correlation. Lastly, most input parameters have a negative correlation coefficient with porosity except for GR and RS, which exhibited a positive correlation with porosity. The result of regression analysis and cross plots in Well 1 of each of the predicted outputs from MLR and the actual output (XRD from core logs) are presented in Fig. 4.1.

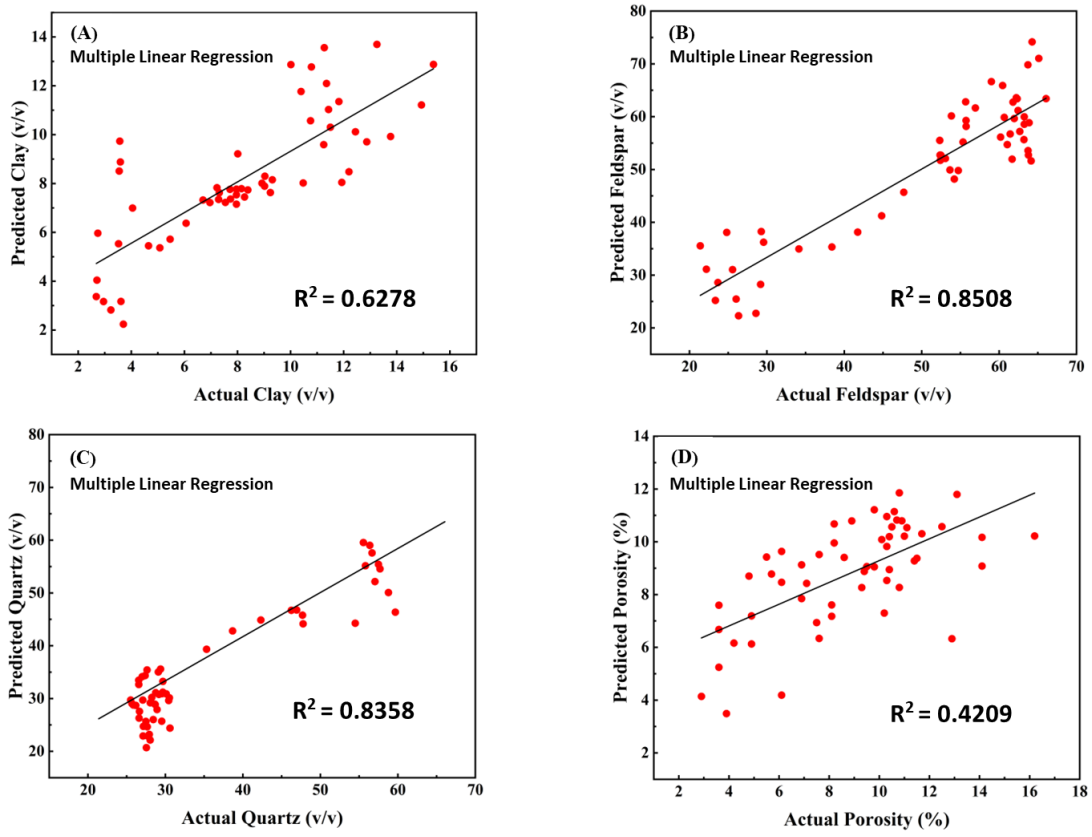


Fig. 4.1 Coefficient of determination (R^2) for the overall datasets of the Multiple Linear Regression analyses in Well 1, volumes of (A) clay minerals, (B) feldspar minerals, (C) quartz minerals, and (D) porosity

It can be observed that the R^2 of clay, feldspar, and quartz mineral contents are above 60% except for porosity which shows a relatively low R^2 of 42% which suggests the model's effectiveness in capturing the general porosity trend while also indicating a need for improved model refinement to explain the variance in the data fully Fig. 4.1.

4.3.2 Equation generated from the MLR method in Well 1

The equations (4.1) - (4.4) were empirically obtained and integrated into the logging software for further interpretation.

$$\text{Predicted Vol_Clay} = 36.001 - 0.284\text{CN} + 0.031\text{GR} - 10.385\text{PE} \quad (4.1)$$

$$\text{Predicted Vol_Feldspar} = 849.548 - 4.039\text{AC} - 215.18\text{DEN} - 0.15\text{GR} + 16.3\text{PE} - 0.868\text{RD} \quad (4.2)$$

$$\text{Predicted Vol_Quartz} = 567.867 + 3.356\text{AC} + 155.241\text{DEN} + 0.132\text{GR} - 15.105\text{PE} + 0.77\text{RD} \quad (4.3)$$

$$\text{Predicted Porosity} = -10.378 + 0.031\text{AC} + 2.749\text{DEN} + 0.300\text{CN} \quad (4.4)$$

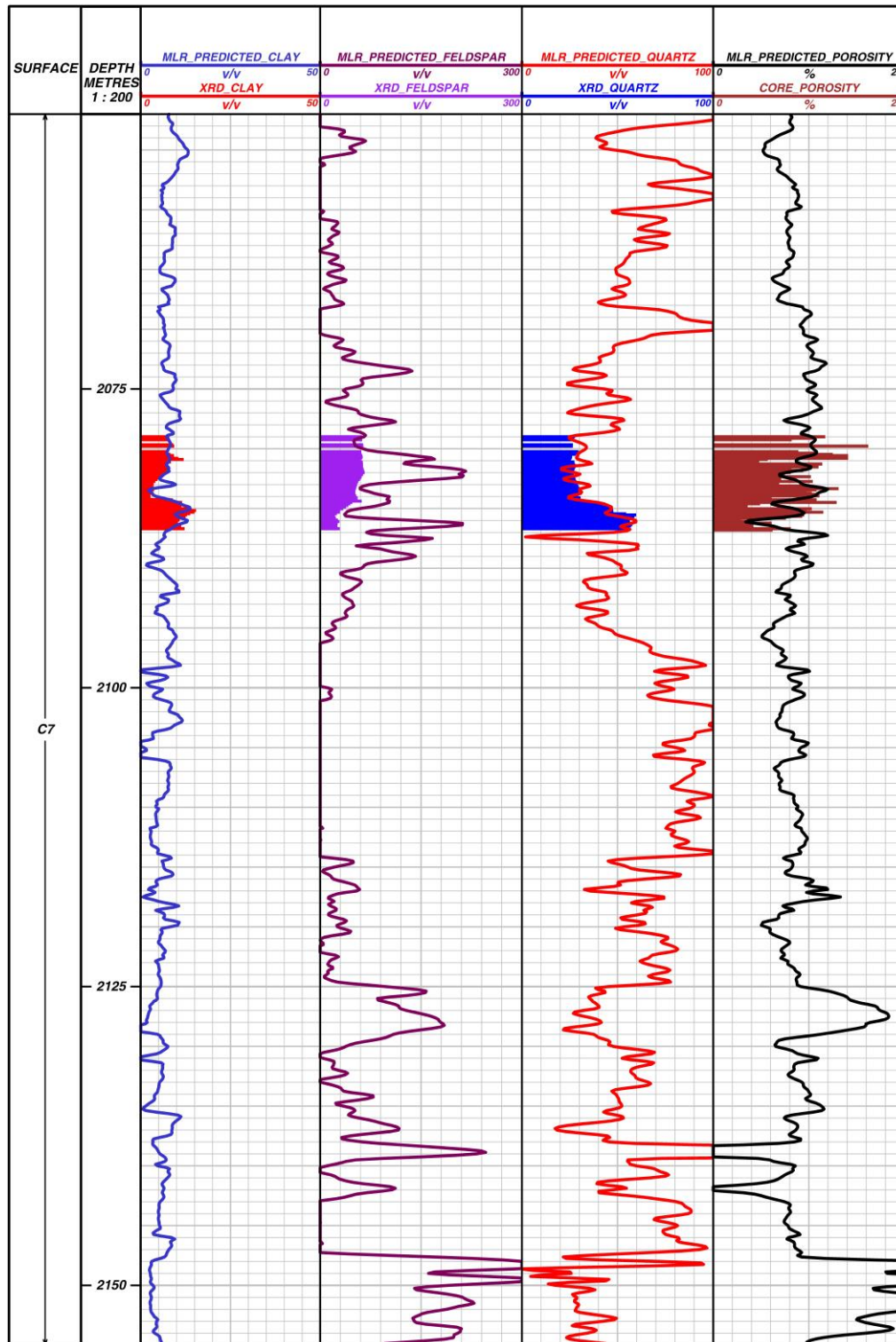


Fig. 4.2 Result of XRD analysis (in red, magenta, blue, and brown color stick) and predicted values (in blue, brown, red, and black curves) of porosity and mineral composition by MLR method in Well 1.

The mineral content (clay, feldspar, quartz) and porosity of well 1 using the MLR method is presented in Fig. 4.2.

4.3.3 Analysis of Mineral Content and Porosity in Well 1 Using MLR Method

Clay Content Analysis

The clay mineral content on track 1, predicted by the Multiple Linear Regression (MLR) method, showed a consistent trend across depths 2052m - 2078m. Near the depth range where XRD clay was measured (2078m - 2086m), the predicted clay content agrees closely with the XRD clay obtained with a high correlation coefficient R^2 of 0.76. However, slight discrepancies were noted at specific intervals, such as 2085m - 2086m, where XRD clay indicated higher clay content than the MLR prediction. Beyond this, fluctuations in predicted clay content were observed, with significant decreases marking the termination of the Formation G reservoir at 2148m-2154m.

Feldspar Content Analysis

The feldspar content on track 2 displayed intermittent presence between 2052m - 2154m, with notable gaps in prediction at several depths, including 2056m-2060m and 2096-2114m, where no predictions were observed. The MLR prediction model overestimated feldspar content when compared to XRD feldspar in the measured interval where the core was obtained (2078m - 2086m), but it has a high correlation coefficient (R^2 of 0.85). Significant spikes in feldspar content were noted at depths between 2147m-2154m, marking the reservoir's termination.

Quartz Content Analysis

Predicted quartz content on track 3 demonstrated significant variability within depths 2052m - 2078m. High fluctuations were noted, reflecting geological heterogeneity or the sensitivity of the MLR model. Between 2078m to 2086m, the predicted quartz content agrees strongly with XRD quartz as it has a high correlation coefficient (R^2 of 0.84). A sharp increase in quartz content was recorded between 2094m and 2124m, followed by a decrease at 2130m and a significant increase at 2148m. These variations indicate dynamic geological processes or depositional changes in the formation.

Porosity Analysis

Porosity predictions on track 4 revealed a consistent trend from depths 2052m - 2115m, aligning closely with core porosity within the 2078m - 2086m interval. Despite having a low correlation coefficient (R^2 of 0.42), good agreement with slight underestimations between core porosity and the predicted porosity was observed. Variations in the porosity zone were noted at depths 2118m - 2125m and 2139m - 2142m, reflecting dynamic reservoir properties. High porosity zones within (2070m - 2095m, 2124m - 2138m, and 2147m - 2154m) aligned with areas of high TOC, as determined by Passey's method, suggesting productive intervals with favorable reservoir quality.

4.4 Chapter Conclusions

Based on the analysis of the mineral content and porosity of Well 1, the MLR model provided strong prediction for clay content with a correlation coefficient R^2 of 0.76, indicating a good fit for most of the formation except for slight discrepancies observed at certain depths (2085m to 2086m). The predicted clay content trend suggests mineralogy variations across the depth profile. Moreover, the predicted feldspar content showed intervals of absence and presence with notable disparities from core data, particularly an overestimation of feldspar content within the depth of XRD-derived feldspar content (correlation coefficient R^2 of 0.85). MLR, in this case, reflects some limitations in predicting capability or variations in geological conditions, as can be seen on track 2 in Fig. 4.2.

Meanwhile, the prediction curve for quartz fluctuated sufficiently, indicating high and low content across different intervals. The predicted quartz is in very good agreement with the XRD quartz having a correlation coefficient of R^2 0.84, demonstrating the model's applicability despite geological heterogeneity. Furthermore, MLR models for porosity predictions were consistent with the core measurement over a broad depth range despite a low correlation coefficient R^2 (0.42), which suggests the model's effectiveness in capturing the general porosity trend while also indicating a need for improved model refinement to explain the variance in the data fully.

Chapter 5 Evaluation of Mineral Composition and Porosity by Machine Learning Approach

5.1 Introduction

In the rapidly evolving area of reservoir characterization, machine learning has emerged as a game-changing tool that provides sophisticated insights that traditional methods often disregard (Mustafa, 2022). Shale formations, with their inherent intricacy and variability, pose a significant challenge for precise mineralogical characterization, which is crucial in determining the quality and potential for reservoir extraction. Two widely accepted machine learning techniques, the Gradient Boost and Extreme Gradient Boost (XGBoost), were used to predict the clay, feldspar, and quartz content, including the porosity of the organic-rich Formation G shale at Ordos Basin.

Using machine learning techniques, GR, AC, NPHI, DEN, RD, and mineralogy logs for about 335 feet in Well 1 thick shale intervals, carrying 102 data points were used to develop the machine learning predictive model for clay, feldspar, and quartz mineral content including the porosity in Formation G Shale formation.

5.2 Methodology

The proposed workflow for mineralogy and porosity prediction with Gradient Boost and Extreme Gradient Boost (XGBoost) for training is presented in Fig. 1.3. The proposed workflow for mineralogy and porosity prediction involves using Gradient Boost and XGBoost models trained on 70% of the original data, comprising conventional well logs (NPHI, DEN, AC, RD, GR) and XRD results for clay, quartz, feldspar, and porosity. The remaining 30% of the data was utilized for blind testing to validate the model's reliability. These trained models were then applied to predict mineral weight fractions at unmeasured depth intervals.

5.3 Results

The statistics of all the input parameters and each output parameter (XRD data) are shown in Supplementary Information in Appendix II (Table 5.1 - Table 5.4) and used for data preprocessing, depth calibration, outlier removal, and normalization when building the prediction model for Gradient Boost and XGBoost. The general approach for building the model will be discussed before presenting each model's performance.

The shape of the empirical probability density function can be described by kurtosis and skewness. As shown in Appendix I (Table 4.2) of supplementary information, Kurtosis

values of all input parameters indicate that data has light tails with minimum chances of having data outliers. A mineral content prediction regression model was developed using the preprocessed dataset, where well-logging parameters were used as input features while the mineral content and porosity were used as the output features. After data preprocessing, a representative subset of data is needed. The re-partition of data points includes a training set, a validation set, and a test set. The training set consists of the range of observation algorithms used to recognize the interaction between the inputs. The validation set is used to adjust the hypo-parameters of the trained model. The test set is defined to evaluate the validated model using a blind prediction of new data.

Data points for all input parameters showed a platy kurtosis distribution, representing good-quality data. These data points lack extreme values, with low and broad peaks. Skewness values of input and output parameters indicate that data is not highly skewed, which evidences data reliability. NPHI, clay, and porosity data exhibited a symmetrical statistical distribution. Moderate skewness was observed for RILD, feldspar, and quartz contents data points.

AC, GR, and DEN data points are negatively skewed, indicating that data points are dominated by values higher than the mean value. NPHI, clay, and porosity data contain data points lower than the mean value, as reflected by its positive skewness. Skewness and kurtosis features for the prediction model of the input features and each output feature of the ML are shown in Appendix II from Table 5.1-5.4 of supplementary information. Box plots were also visualized to observe outliers, identify skewness, and understand the feature engineering attribute shown in (Appendix IV, Fig. 5.5 of supplementary information).

The visualized Pearson's correlation coefficient of all the input features and their output features to understand the relationship between variables is presented in (Appendix V, Fig. 5.6 of the supplementary information). The Heat Map assists in feature selection by highlighting variables strongly correlated with the target variables. Highly correlated features may provide redundant information, and selecting only one of them can simplify the model without sacrificing predictive accuracy.

Subsequently, four models were developed for the prediction of the clay model, feldspar model, quartz model minerals content, and porosity. Seventy percent (70%) of the sample data was selected as the training data set to train the model, and 30% was used as the testing data set to test the model's effectiveness. The sample data set was trained using the 10-fold cross-validation method to generate five independently and identically

distributed training and test data sets, which were built separately, allowing the models to be adequately trained and validated to obtain the best inversion accuracy as shown in the supplementary information (Appendix V, Fig. 5.6).

Evaluation metrics are quantitative metrics used to describe how well or poorly a model works for the same data fed into different algorithmic models or different data fed into the same algorithmic model, including regression and classification evaluation metrics. Regression evaluation metrics: Regression models used MAE, RMSE, and R^2 to evaluate performance. MAE is the mean absolute difference between the predicted and actual values, representing the average magnitude of the error in a set of predictions. RMSE is the root mean square of the difference between the predicted and actual values, which indicates the closeness of the observed data points to the model's predicted values. R^2 reflects the overall degree of fitting of the model. The lower the MAE and RMSE values, the better the model's predictive performance. In contrast, the higher the R^2 value, the better the model's prediction performance. These three metrics are defined as follows equation

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (5.1)$$

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (5.2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5.3)$$

where y and $\hat{y}$ are the observed values and model-predicted values from the dataset, $\bar{y}$ is the mean of the observed values, and n is the number of samples.

After discussing the general data preprocessing and processing stages, two widely used machine learning approaches will then be implemented to build a model that will be used to make predictions in Well 1. The presentation of each performance of the machine learning method is as follows.

5.3.1 Gradient Boost Model Performance

The performance of gradient-boosting algorithms is remarkable due to their capacity for detecting patterns and trends that can be difficult for human analysts to discern. The gradient boost technique will further be used to develop a prediction model to predict the mineral composition and porosity in Well 1. Appendix I (Table 4.1) displays the

correlation coefficient matrix of input parameters with output parameters for Well 1 to determine the parameters for training the data to avoid redundancy.

After determining the correlation coefficient of the input and output features, the correlation coefficient for the overall data set, the Mean Absolute Error (MAE), and the Root Mean Squared Error are then determined. The cross plot is carried out to know the model's performance in predicting the mineral composition in the shale reservoir. The square of the correlation coefficient (R^2) values obtained from the training data set were 0.999, 0.998, 0.999, and 0.901, correlation of determination (R^2) values for the test data set were; 0.862, 0.925, 0.853, and 0.606 with an overall coefficient of determination (R^2) of 0.952, 0.975, 0.950 and 0.783, the MAE for training is 0.08, 0.29, 0.11 and 1.02 while MAE for testing is 1.15, 2.63, 1.87 and 2.25; RMSE of 1.28, 2.94, 2.47 and 2.49 was obtained from the Gradient Boost model that has been built to predict the mineral composition of clay, feldspar, quartz, and porosity. Table 5.1 shows the evaluation metrics for the gradient boost regression model while Fig. 5.1 - Fig. 5.4 shows the cross-plots for the training, testing, and the overall dataset for the predicted model.

Table 5.1 Statistics of evaluation metrics for Gradient Boost regression model.

Parameters	Training MAE	Testing MAE	RMSE	Overall R^2
Clay Volume	0.08	1.15	1.28	0.952
Feldspar Volume	0.29	2.63	2.94	0.975
Quartz Volume	0.11	1.87	2.47	0.950
Porosity	1.02	2.25	2.49	0.783

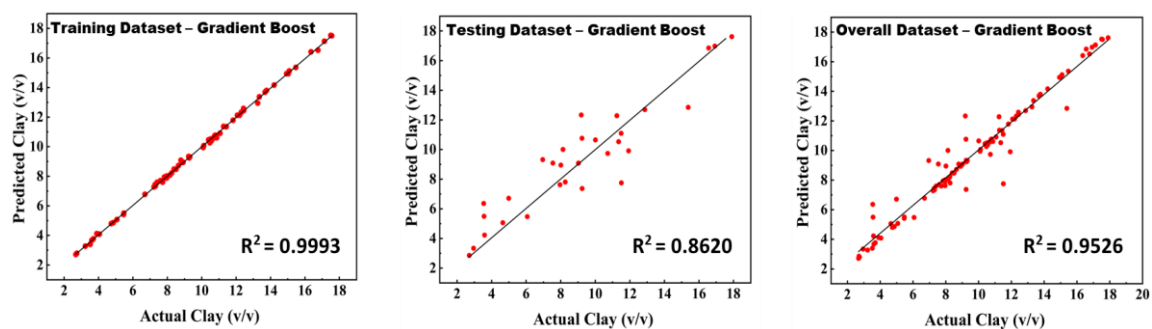


Fig. 5.1: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for clay volume

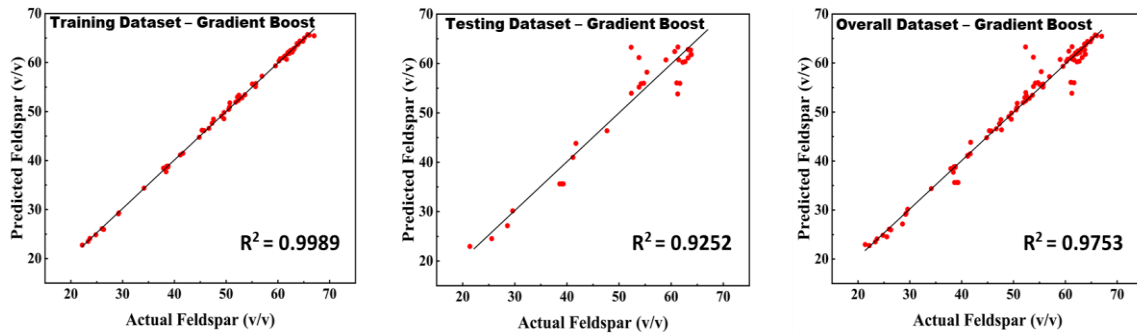


Fig. 5.2: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for feldspar volume

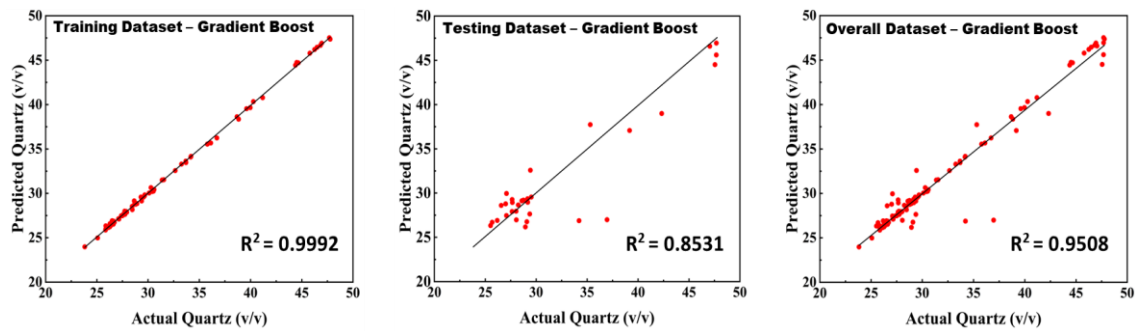


Fig. 5.3: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for quartz volume

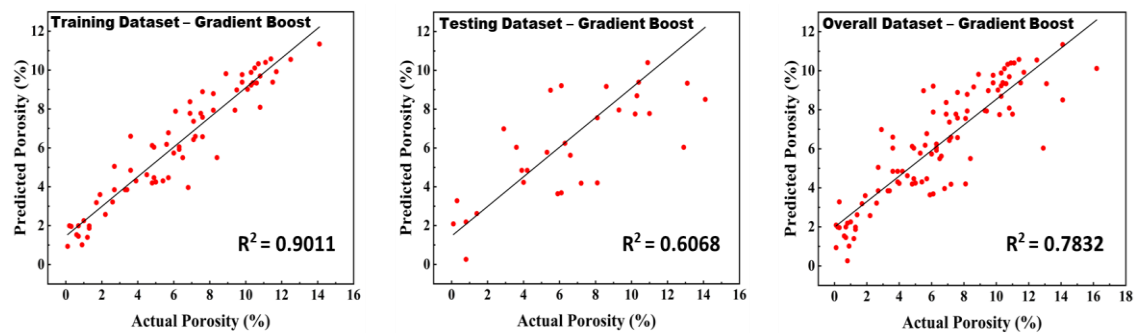


Fig. 5.4 The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for porosity

The feature importance of each input variable and permutation feature importance was also obtained (Appendix VI, Fig. 5.12 - 5.15) for the Gradient Boost. The permutation feature importance is defined as the decrease in a model score when a single feature value is randomly shuffled (Tahmasebi, 2017). Permutation feature importance evaluates input and output features by calculating the drop in the model score.

The bigger the drop in the model score, the more impact the input feature has on the model and the higher it will rank. Permutation importance can be obtained by importing the "permutation_importance" library from "sklearn.inspection" and passing in the trained model (in this case, "gb"), testing set (in this case, "X_test, y_test"), n_repeats (in this case, 10 was used), and random_state (number of seeds).

Following the development of the predicted model by gradient boost, similar conventional well logging data used to build the model was obtained from Well 1. These well log data obtained from Well 1 were trained into the model. The predictive outputs were subsequently uploaded to the petrophysical software to generate synthetic wireline logs. These synthetic logs allowed for a direct comparison with the actual core data, enabling a more nuanced interpretation of the reservoir's characteristics. The interpreted logs in Well 1 are discussed subsequently.

The Gradient-Boosted model was diligently calibrated using a training dataset derived from Well 1, encompassing a comprehensive suite of AC, GR, RILD, DEN, and NPHI wireline logs. Upon validation, the model exhibited robust predictive capabilities, with the following performance metrics: mean absolute error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2), as in Table 5.1.

5.3.2 Gradient Boost Results in Well 1

The interpreted log from the predicted values and core log obtained from the integration of the well log data of Well 1 into the gradient boost algorithm model is presented in Fig. 5.5. From left to right, track 1 represents the surface interval; Track 2 represents the measured depth; Track 3 compares XRD clay and predicted clay; Track 4 presents XRD feldspar and predicted feldspar; and Track 5 shows XRD quartz and predicted quartz. Track 6 also presents the comparison of core porosity and predicted porosity.

It can also be observed that the predicted mineral content and porosity were in good agreement with the core data when it approached the depth of 2390m - 2396m, as shown in Fig. 5.5 except for the quartz content and porosity. The predicted quartz content is underestimated relative to the XRD quartz between the depths of 2085m - 2086m and the predicted porosity between the depths of 2079m - 2081m, respectively. As observed on the interpreted log, the mineral contents and porosity were uniform across all depths until the depth of 2416m - 2431m, when a subsequent increase was observed in the predicted mineral content.

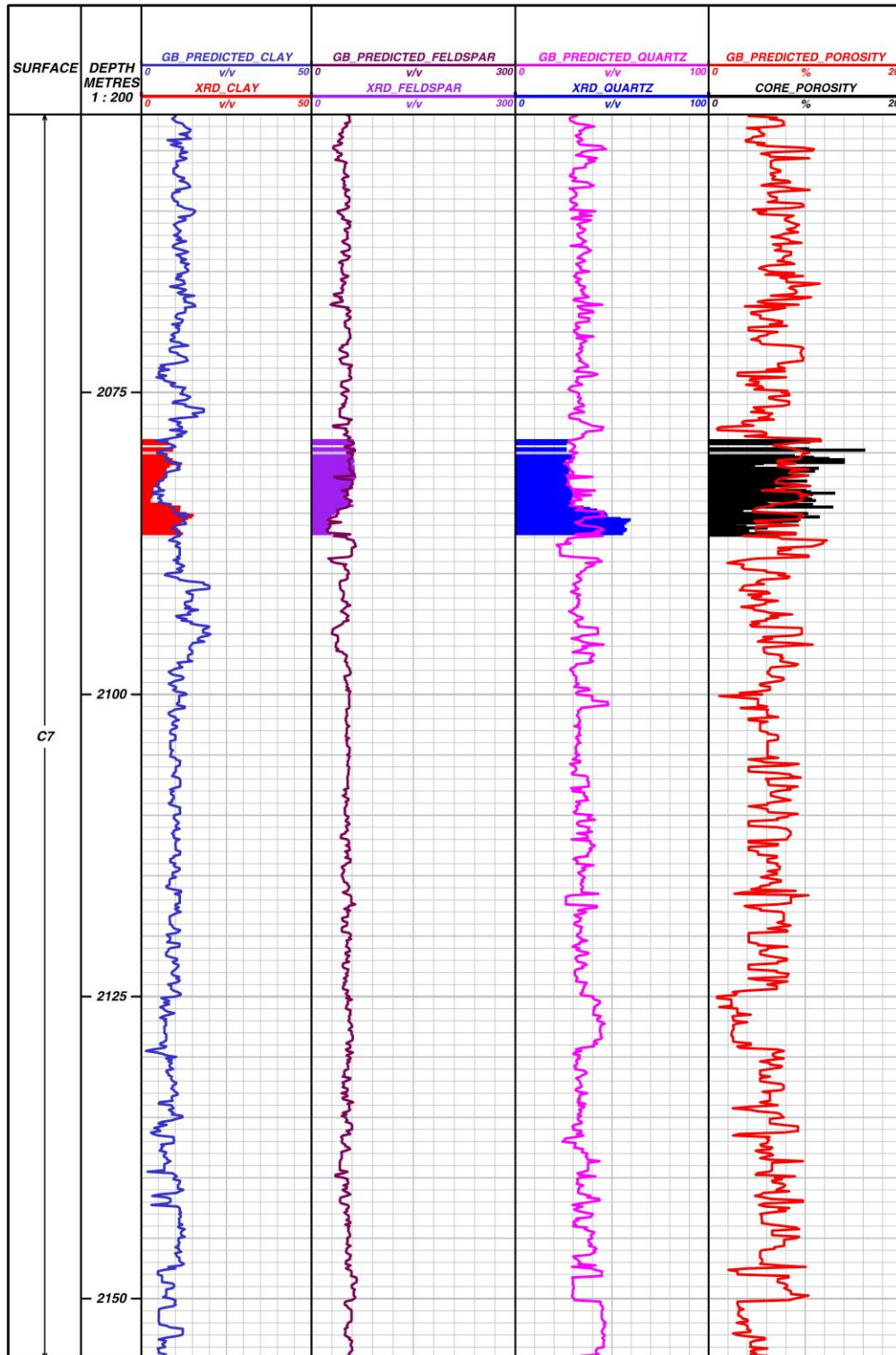


Fig. 5.5: Results of XRD analysis (in red, magenta, blue, and black color stick) and predicted values (in blue, brown, magenta, and red curves) for mineral composition and porosity by Gradient Boost in Well 1.

5.3.3 Extreme Gradient (XGBoost) Model Performance

The XGBoost algorithm is an advanced ensemble learning method that excels at managing complex data spaces. It is particularly well-suited for the challenges of subsurface data, enabling us to gain a comprehensive understanding of the reservoir's distribution. This is crucial for making strategic decisions in exploration and production operations. This analysis demonstrates the effectiveness of machine learning models in capturing intricate patterns within the geophysical data, revealing the characteristics of the shale reservoir under investigation. XGBoost technique will further be used to develop a prediction model to predict the mineral composition and porosity in Well 1. Appendix I (Table 4.1) of the supplementary information shows the correlation coefficient matrix of input parameters with output parameters for Well 1 to determine the parameters for training the data to avoid redundancy.

After determining the coefficient of the input and output features, the correlation coefficient for the overall data set, the Mean Absolute Error (MAE), and the Root Mean Squared Error are then determined. The cross plot was carried out to determine the model's performance in predicting the mineral composition in the shale reservoir. The square of the correlation coefficient (R^2) values obtained from the training data set was 0.983, 0.986, 0.921, and 0.934, the square of the correlation coefficient (R^2) values for the test data set; were 0.745, 0.741, 0.682, and 0.488 with an overall coefficient of determination (R^2) of 0.888, 0.908, 0.779 and 0.753; MAE of training is 0.50, 0.91, 1.18 and 0.84 while MAE for testing is 1.80, 5.29, 3.23 and 2.39; RMSE of 1.79, 5.42, 4.03 and 2.59 from XGBoost model that has been built to predict the mineral composition of clay, feldspar, quartz, and porosity shows the evaluation metrics for the XGBoost boost regression model are displayed in Table 5.2 while the cross plots for the training, testing, and the overall dataset for the predicted model are presented in Fig. 5.6-5.20.

Table 5.2 Statistics of the evaluation metrics for the XGBoost regression model

Parameters	Training MAE	Testing MAE	RMSE	Overall R^2
Clay volume	0.50	1.80	1.79	0.888
Feldspar volume	0.91	5.29	5.42	0.908
Quartz volume	1.18	3.23	4.03	0.779
Porosity	0.84	2.39	2.59	0.753

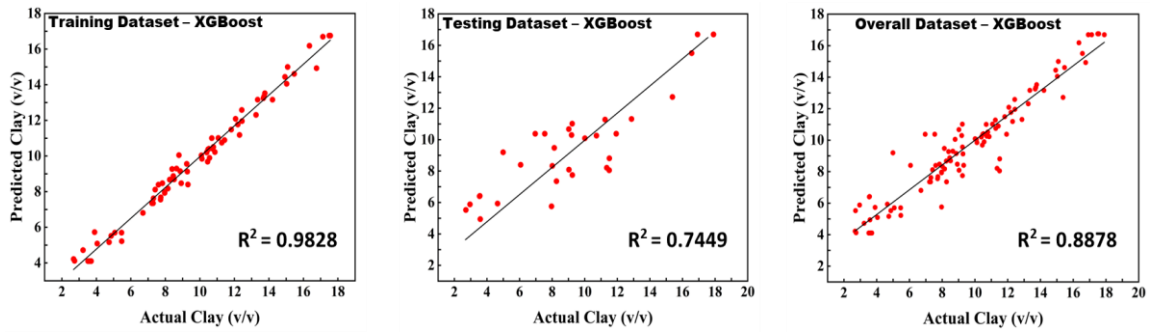


Fig. 5.6: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for clay volume.

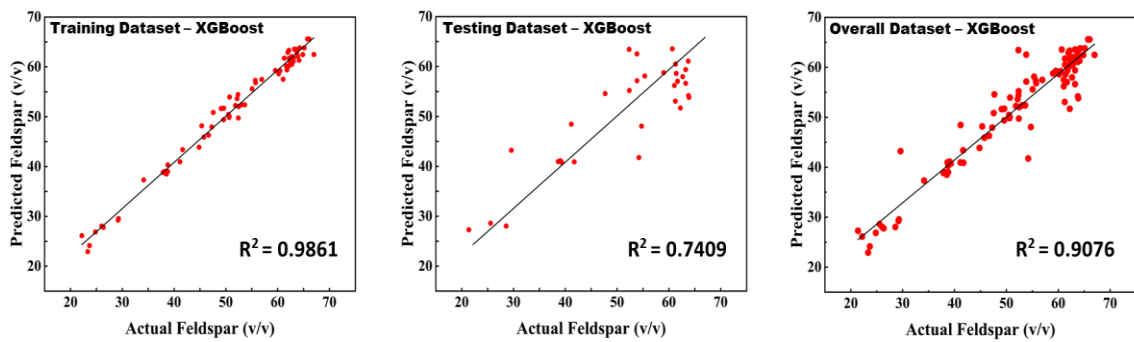


Fig. 5.7: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for feldspar volume

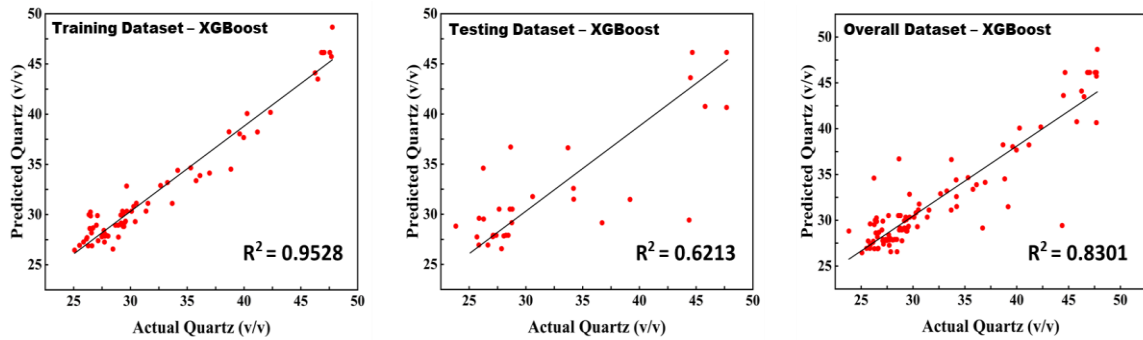


Fig. 5.8: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for quartz volume.

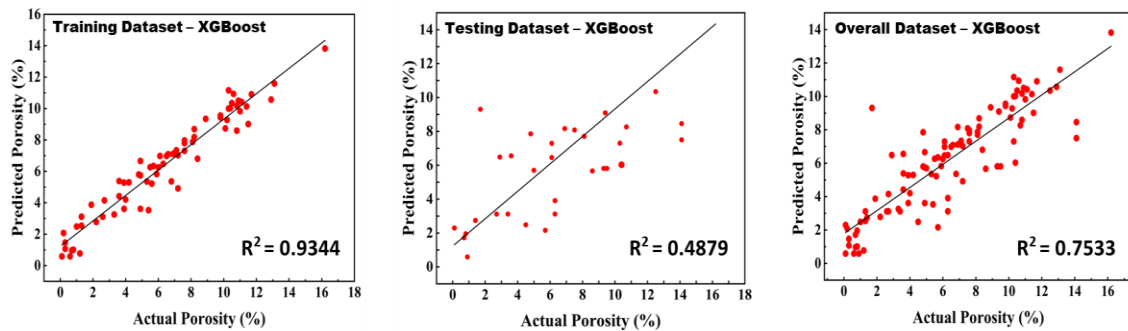


Fig. 5.9: The cross-plot of Predicted Training and Testing and Overall Data against Actual Data for porosity.

The feature importance of each input variable and permutation feature importance was also obtained. The importance of permutation tends to illustrate how essential a feature is for a model (Appendix VII, Fig. 5.21-5.24 of supplementary information).

Following the development of the predicted model using XGBoost, similar conventional well log data used to build the model will be obtained in Well 1. This well-log data was employed to train the model to obtain a predictive output. The predictive outputs were subsequently imported to the petrophysical software to generate synthetic wireline logs. These synthetic logs allowed for a direct comparison with the actual core data, enabling a clearer interpretation of the reservoir's characteristics. The XGBoost model encompassing a comprehensive suite of AC, GR, RILD, DEN, and NPHI wireline logs was calibrated with a training dataset derived from Well 1. Upon validation, the model exhibited robust predictive capabilities, with the following performance metrics; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), as shown in Table 5.2.

5.3.4 XGBoost Result in Well 1

Fig. 5.10 presents the well-log interpretation of the predicted values and core log obtained from the well log data of Well 1 using the XGBoost algorithm model. The well log is divided into six tracks, each representing different information.

Track 3 displays a consistent trend in the predicted curve of the clay content, while track 4 indicates the predicted curve of the feldspar volume. Track 5 represents the predicted curve of the quartz volume, which underestimated the measured XRD at certain depths. Track 6 shows the predicted porosity curve, which consistently increased and decreased through the depth.

The well-log analysis shows that the predicted values agreed with the actual XRD data, except for specific depths where the predicted curves underestimated the actual XRD data.

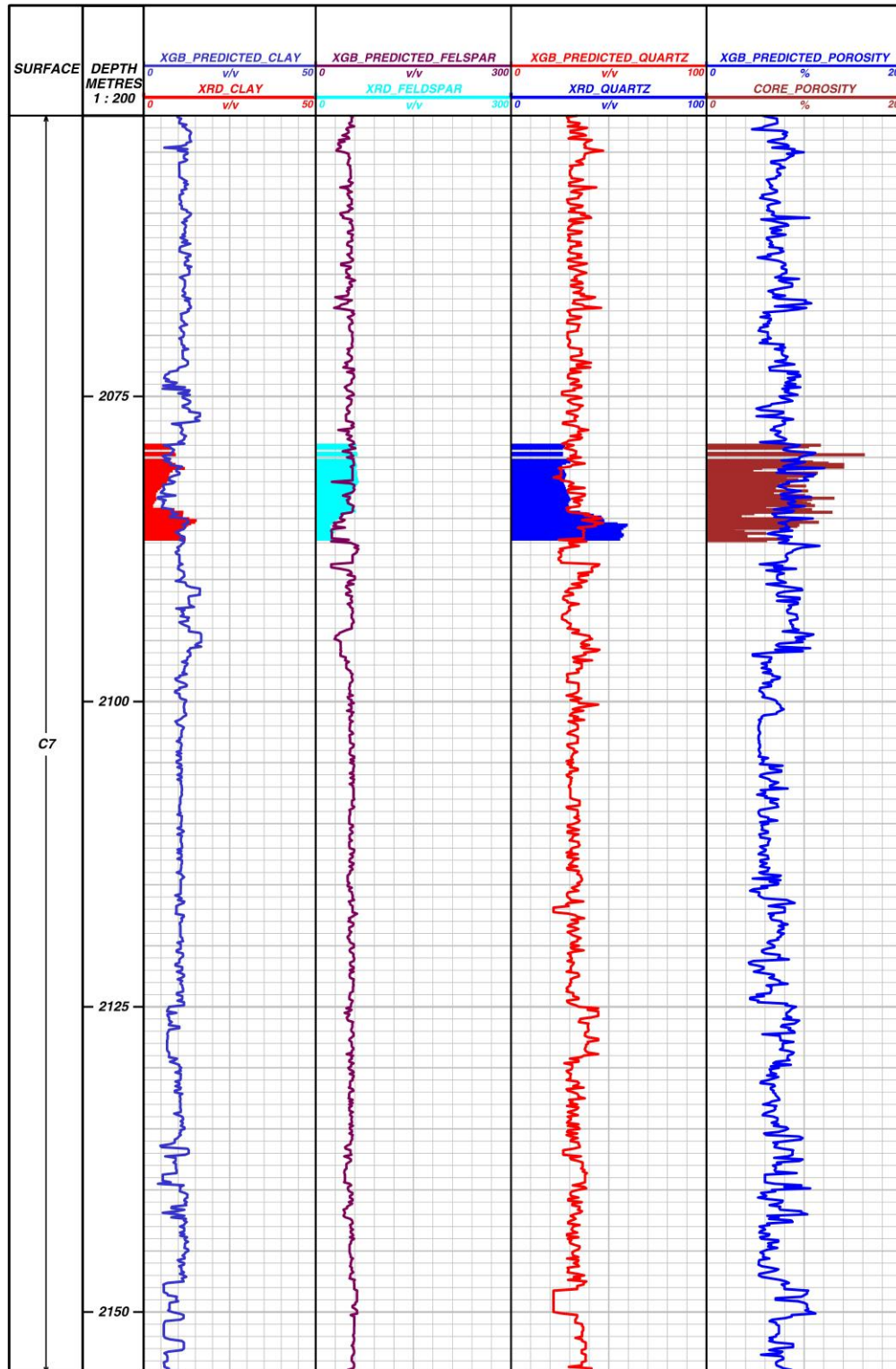


Fig. 5.10: The result of XRD analysis (in red, cyan, blue, and brown color stick) and predicted (in violet, earth, red, and blue curves) mineral composition and porosity using XGBoost in Well 1.

5.4 Performance and Comparative Analysis of MLR, Gradient Boost, and XG Boost in Evaluating Mineral Composition and Porosity.

This study demonstrates the effectiveness and accuracy of predictive models, including MLR, Gradient Boost, and XGBoost, for identifying mineral composition in shale reservoirs based on specific input parameters as these models achieve excellent accuracy with minimal errors across various datasets, such as training, testing, validation, and overall data. For example, the overall accuracy of the MLR model in Well 1 is shown in Fig. 5.11 respectively.

The Gradient Boost model surpassed both the XGBoost and MLR models in predicting the contents of clay, feldspar, quartz contents, and porosity. Furthermore, the Gradient Boost model offers a cost-effective and practical method for estimating the content of clay, feldspar, quartz, and porosity through available well logs. Comparative analysis of all applied methods in Well 1

In Well 1, the mineral composition and porosity were analyzed using various methods, and the results are depicted in Fig. 5.11. The predicted curves for the clay, feldspar, quartz contents, and porosity using MLR, Gradient Boost, and XGBoost methods agreed with each other, except for some fluctuations and discontinuities in some depth intervals. The predicted curves for the feldspar content were accurately predicted by the Gradient Boost and XGBoost methods but not by the MLR method. On the other hand, the predicted curves for the quartz content agreed with the measured XRD quartz. An increase in quartz content was noticeable at the depth where a decrease in the feldspar content was observed. Overall, the results are auspicious and provide valuable insights into the mineral composition and porosity in Well 1.

5.5 Chapter Conclusions

In summary, MLR, Gradient Boost, and XGBoost prediction models were developed to predict the mineral composition in Well 1. MLR, Gradient Boost, and XGBoost models showed strong predictions for mineral composition and core porosity. Both wells had excellent agreement with each other, highlighting the models' robustness. The Gradient Boost performed exceptionally well, suggesting its effectiveness in predicting mineral composition and porosity. The error comparison of training and testing results for clay, feldspar, quartz contents, and porosity prediction model based on the square of the correlation coefficient R^2 of the Gradient Boost and XGBoost is presented in Appendix VIII (Fig. 5.27) of supplementary information.

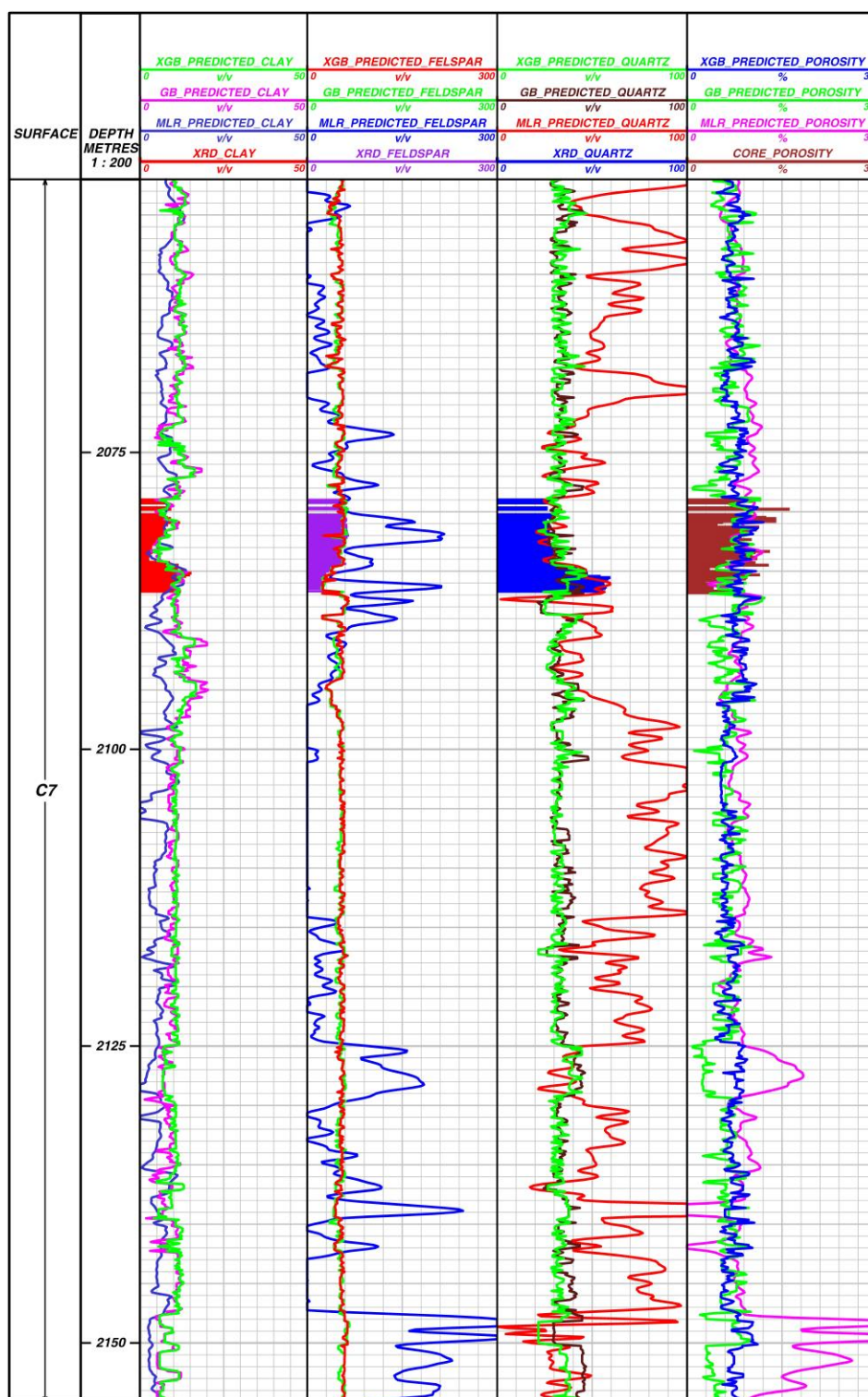


Fig. 5.11: XRD analysis for estimating the mineral composition and core porosity (in red, purple, blue, and magenta curves) and the curve of predicted mineral composition using MLR, Gradient Boost, and XGBoost method in Well 1.

Chapter 6 Conclusions and Recommendations

6.1 Conclusions

A comprehensive evaluation, assessment, and characterization of TOC, mineral composition, and porosity of Formation G shale Formation Gas performed from well logs. From this study, the following conclusions can be made. TOC analysis results for the Well showed a good agreement between the estimated organic content and the measured TOC from the core samples at the depth where it was measured. It was observed that high TOC zones conformed with high porosity zones. This is probably attributed to the organic matter (kerogen) undergoing thermal maturation that generates hydrocarbon and releases gases, thereby increasing pore spaces within the shale reservoir. Moreover, using ML models like Gradient Boost, Adaptive Gradient Boost, Random Forest, and Stacking techniques to evaluate TOC, ML models such as Gradient Boost and XG Boost to predict mineralogy and porosity generated results that were in good agreement with borehole data (core data and conventional well logging data).

The Gradient Boost mineral composition evaluation model exhibited superior identification metrics than the MLR and XGBoost evaluation models with lower error and higher R^2 . However, multiple linear regression was observed to overestimate the mineral content for feldspar in well 1 and slightly underestimate the actual porosity content in the same well. Based on Root Mean Square (RMSE) and Mean Arithmetic Error (MAE) analysis, the Machine Learning model-based predicted models gave more reliable and accurate results, which confirms the effectiveness and reliability of the proposed methods to predict mineral composition and porosity in Well 1 from well logging measurements.

This study has shown that predictive models could be cost-effective for determining TOC, clay, feldspar, quartz content, and petrophysical properties like porosity, reiterating this research's significance. Moreover, this two-pronged strategy of combining conventional well log interpretation with MLR, ML, and stacking technique represents a progressive step in the forecasting of reservoir properties facilitating better-informed decision-making for the exploration and development of unconventional hydrocarbon resources that would be very helpful in targeting and optimizing hydraulic fracturing in high-productive shale reservoir zones.

6.2 Future Research Prospectives

Considering the outcomes and results obtained from this study and the potential implications for future research and practical applications, it is suggested that Passey's MLR, ML, and Stacking methods should be extended across different shale formations to validate their adaptability to validating geological contexts. Also, integrating additional well logs should be explored to refine the TOC estimation provided by Passey's and MLR method and potentially incorporate machine learning algorithms for enhanced precision. Besides, the method's performance should be compared against other TOC estimation methods in shale plays to ascertain the most effective approach.

Gradient Boosting, Adaptive Boosting, Random Forest, XGBoost models, stacking, bagging, and boosting should be applied to a large dataset to test further and improve the robustness of TOC and mineral predictions. Real-time data analysis using machine learning to facilitate dynamic reservoir characterization and management requires further study.

Lastly, interdisciplinary studies combining geological, geochemical, and engineering data to exploit machine learning's predictive capabilities in resource assessment should be considered. Large datasets should be available, as machine learning techniques require appropriate data to build a prediction model.

Appendix Own Results Summary

Scientific Question Solved

This research addressed the challenge of accurately characterizing shale reservoirs by concentrating on estimating and evaluating Total Organic Carbon (TOC), mineral composition, and porosity using well logs. The study sought to bridge the divide between traditional evaluation methods and advanced machine learning techniques, thereby improving prediction accuracy and fostering a more comprehensive understanding of shale reservoirs. A key investigation centered on the effectiveness of integrating multiple predictive models to enhance the assessment of shale reservoirs in Formation G of the Ordos Basin.

Methods Personally Applied

The following methods were applied in this research.

1. Development of Predictive Models:

- a. I created and applied predictive frameworks utilizing Passey's $\Delta\log R$ approach, Multiple Linear Regression (MLR), and advanced machine learning methods such as Gradient Boost, Adaptive Gradient Boost, Random Forest, and the Stacking Ensemble method.
- b. Conventional well log data (e.g., gamma ray, resistivity, density, and neutron logs) and core data were collected and analyzed to establish robust models.

2. Data Integration and Interpretation:

- a. I incorporated the model outputs into petrophysical software for detailed visualization of lithological and petrophysical characteristics.
- b. Machine learning predictions were combined with well-log responses to identify high TOC zones and delineate brittle and ductile layers within the shale formation.

3. Model Validation and Analysis:

- a. Statistical metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 (coefficient of determination) were applied to assess and compare model performance.
- b. Sensitivity analyses were also conducted to evaluate the influence of input variables on predictive accuracy.

Results Obtained

1. TOC Evaluation:

- a. The TOC successfully demonstrated strong correlations between predicted TOC values and core-measured data using Passey's method, MLR, and machine learning techniques.
- b. In this research, the Gradient Boost models and Stacking technique showed superior performance in the training and testing phases and, as such, outperformed other methods, achieving higher prediction accuracy with minimal errors, particularly in organic-rich zones.

2. Mineral Composition and Porosity Prediction:

- a. This work successfully predicted the mineralogical composition (clay, feldspar, and quartz) and porosity of Formation G shale using multiple linear regression and machine learning techniques.
- b. Gradient Boost models exhibited exceptional training and testing phase results, providing reliable predictions for reservoir characterization.

3. Integration and Workflow Optimization:

- a. Integrating the predictive outputs into petrophysical software to create depth-matched wireline log interpretations, facilitating more precise reservoir characterization.
- b. There is a reduced dependency on traditional core-based methods, offering cost-effective and scalable alternatives through data-driven techniques.

4. Practical Contributions to Reservoir Management:

- a. This research identified zones with high TOC and favorable brittleness characteristics, which aid in selecting optimal hydraulic fracturing targets.
- b. There is an improved understanding of heterogeneity in shale reservoirs, leading to better-informed exploration and production strategies.

In conclusion, this research demonstrates the transformative potential of integrating traditional geophysical methods with machine learning models, providing a framework for enhanced evaluation and management of unconventional reservoirs globally.

References

- Abordan, A., & Szabó, N. (2018). Particle swarm optimization assisted factor analysis for shale volume estimation in groundwater formations. *Geosciences and Engineering; Miskolc*, 6(9), 87–97.
- Abouelresh, M. O., & Slatt, R. M. (2012). Lithofacies and sequence stratigraphy of the Barnett Shale in east-central Fort Worth Basin, Texas. *AAPG Bulletin*, 96(1), 1–22.
- Abudeif, A. M., Attia, M. M., & Radwan, A. E. (2016). Petrophysical and petrographic evaluation of Sidri Member of Belayim Formation, Badri field, Gulf of Suez, Egypt. *Journal of African Earth Sciences*, 115, 108–120.
- Amosu, A., & Sun, Y. (2021). Identification of thermally mature total organic carbon-rich layers in shale formations using an effective machine-learning approach. *Interpretation*, 9(3), T735–T745.
- Bilgen, S., & Sarıkaya, İ. (2016). New horizon in energy: Shale gas. *Journal of Natural Gas Science and Engineering*, 35, 637–645. <https://doi.org/10.1016/j.jngse.2016.09.014>
- Boz, M. A. (2014). Recycling of Produced and Flowback Water in Oil and Gas Drilling Operations through Hydraulic Fracturing in Texas. *Shale Energy Engr.* 2014, 45–52.
- Chen, S., Zhao, W., Ouyang, Y., Zeng, Q., Yang, Q., Hou, H., Gai, S., Bao, S., & Li, X. (2017). Prediction of sweet spots in shale reservoir based on geophysical well logging and 3D seismic data: A case study of Lower Silurian Longmaxi Formation En W4 block, Sichuan Basin, China. *Energy Exploration & Exploitation*, 35(2), 147–171.
- Clarkson, C. R., Solano, N., Bustin, R. M., Bustin, A. M. M., Chalmers, G. R. L., He, L., Melnichenko, Y. B., Radliński, A. P., & Blach, T. P. (2013). Pore structure characterization of North American shale gas reservoirs using USANS/SANS, gas adsorption, and mercury intrusion. *Fuel*, 103, 606–616.
- Conti J., Holtberg, P., Diefenderfer, J., Napolitano, S., Schaal, A. , Turnure, J. , & Westfall, L. ., (2014). *Annual Energy Outlook 2014 with Projections to 2040*.
- Ding, W., Li, C., Li, C., Xu, C., Jiu, K., Zeng, W., & Wu, L. (2012). Fracture development in shale and its relationship to gas accumulation. *Geoscience Frontiers*, 3(1), 97–105.
- Dong, D., Shi, Z., Guan, Q., Jiang, S., Zhang, M., Zhang, C., Wang, S., Sun, S., Yu, R., Liu, D., Peng, P., & Wang, S. (2018). Progress, challenges and prospects of shale gas exploration in the Wufeng–Longmaxi reservoirs in the Sichuan Basin. *Natural Gas Industry B*, 5(5), 415–424.

- Elmahdy, M., Radwan, A. A., Nabawy, B. S., Abdelmaksoud, A., & Nastavkin, A. V. (2023). Integrated geophysical, petrophysical and petrographical characterization of the carbonate and clastic reservoirs of the Waihapa Field, Taranaki Basin, New Zealand. *Marine and Petroleum Geology*, 151, 106173.
- Fu, J., Li, S., Niu, X., Deng, X., & Zhou, X. (2020). Geological characteristics and exploration of shale oil in Formation G Member of Triassic Yanchang Formation, Ordos Basin, NW China. *Petroleum Exploration and Development*, 47(5), 931–945.
- Gale, J. F. W., Laubach, S. E., Olson, J. E., Eichhuble, P., & Fall, A. (2014). Natural Fractures in shale: A review and new observations. *AAPG Bulletin*, 98(11), 2165–2216.
- Ghahremani, N., & Clapp, L. (2014). Feasibility of Using Brackish Groundwater Desalination Concentrate as Hydraulic Fracturing Fluid in the Eagle Ford Shale. *Shale Energy Engineering 2014*, 23–32.
- Gogoi, T., & Chatterjee, R. (2019). Estimation of petrophysical parameters using seismic inversion and neural network modeling in Upper Assam basin, India. *Geoscience Frontiers*, 10(3), 1113–1124.
- Huang, F., Yang, T., & Yan, W. (2013). Geologic Factors of Formation of Tight Oil and Its Resource Potential in China. *AAPG Annual Convention and Exhibition, Pittsburgh, Pennsylvania*.
- International Energy Agency (IEA). (2017). *Digitalisation and Energy*.
- Jacobi, D., Breig, J., LeCompte, B., Kopal, M., Hursan, G., Mendez, F., Bliven, S., & Longo, J. (2009 October 4). Effective Geochemical and Geomechanical Characterization of Shale Gas Reservoirs from the Wellbore Environment: Caney and the Woodford Shale. *All Days*.
- James W. Schmoker¹. (2002). Resource-assessment perspectives for unconventional gas systems. *AAPG Bulletin*, 86.
- Jarvie, D. M., Hill, R. J., Ruble, T. E., & Pollastro, R. M. (2007). Unconventional shale-gas systems: The Mississippian Barnett Shale of north-central Texas as one model for thermogenic shale-gas assessment. *AAPG Bulletin*, 91(4), 475–499.
- Javadpour, F., Fisher, D., & Unsworth, M. (2007). Nanoscale Gas Flow in Shale Gas Sediments. *Journal of Canadian Petroleum Technology*, 46(10).
- Jiao, F. (2019). Theoretical insights, core technologies and practices concerning “volume development” of shale gas in China. *Natural Gas Industry B*, 6(6), 525–538.

- Jiao, F., Zou, C., & Yang, Z. (2020). Geological theory and exploration & development practice of hydrocarbon accumulation inside continental source kitchens. *Petroleum Exploration and Development*, 47(6), 1147–1159.
- Josh, M., Esteban, L., Delle Piane, C., Sarout, J., Dewhurst, D. N., & Clennell, M. B. (2012). Laboratory characterisation of shale properties. *Journal of Petroleum Science and Engineering*, 88–89, 107–124.
- Kumar, R., Das, B., Chatterjee, R., & Sain, K. (2016). A methodology of porosity estimation from inversion of post-stack seismic data. *Journal of Natural Gas Science and Engineering*, 28, 356–364.
- Li, Y., & Han, Y. (2017, November 7). Decline Curve Analysis for Production Forecasting Based on Machine Learning. *Day 1 Tue, November 07, 2017*.
- Liu, K., Xu, B., Kim, C., & Fu, J. (2021). Well Performance from Numerical Methods to Machine Learning Approach: Applications in Multiple Fractured Shale Reservoirs. *Geofluids*, 2021, 1–13.
- Liu, W., Zhang, G., Cao, J., Zhang, J., & Yu, G. (2019). Combined petrophysics and 3D seismic attributes to predict shale reservoirs favourable areas. *Journal of Geophysics and Engineering*, 16(5), 974–991.
- Ma, L., Zhang, Z., & Meng, W. (2021). Climate-Provenance Effect on the Organic Matter Enrichment of the Formation I Source Rocks in the Central Ordos Basin, China. *Geofluids*, 2021, 1–10.
- Mustafa, A., Tariq, Z., Mahmoud, M., Radwan, A. E., Abdulraheem, A., & Abouelresh, M. O. (2022). Data-driven machine learning approach to predict mineralogy of organic-rich shales: An example from Qusaiba Shale, Rub’ al Khali Basin, Saudi Arabia. *Marine and Petroleum Geology*, 137, 105495.
- Nhabanga, O. J., Ringrose, P. S., & Holt, R. M. (2021). Use of rock-physics analysis of well logs to determine compaction history of Cretaceous shales in the Rovuma basin, Offshore Mozambique. *Geophysical Prospecting*, 69(6), 1282–1294.
- Nicot, J.-P., & Scanlon, B. R. (2012). Water Use for Shale-Gas Production in Texas, U.S. *Environmental Science & Technology*, 46(6), 3580–3586.
- Niu, W., Lu, J., Sun, Y., Liu, H., Cao, X., Zhan, H., & Zhang, J. (2023). A review of the application of data-driven technology in shale gas production evaluation. *Energy Reports*, 10, 213–227.

- Ouenes, A. (2014, February 25). Distribution of Well Performances in Shale Reservoirs and Their Predictions Using the Concept of Shale Capacity. *Day 2 Wed, February 26, 2014*.
- Passey, Q. R., Bohacs, K. M., Esch, W. L., Klimentidis, R., & Sinha, S., (2010, June 8). From Oil-Prone Source Rock to Gas-Producing Shale Reservoir – Geologic and Petrophysical Characterization of Unconventional Shale-Gas Reservoirs. *All Days*.
- Q. R. Passey, S. C., J. (1990). A Practical Model for Organic Richness from Porosity and Resistivity Logs. *AAPG Bulletin*, 74.
- Radwan, A. A., & Nabawy, B. S. (2022). Hydrocarbon prospectivity of the miocene-pliocene clastic reservoirs, Northern Taranaki basin, New Zealand: integration of petrographic and geophysical studies. *Journal of Petroleum Exploration and Production Technology*, 12(7), 1945–1962.
- Radwan, A. E., Wood, D. A., Mahmoud, M., & Tariq, Z. (2022). Gas adsorption and reserve estimation for conventional and unconventional gas resources. *Sustainable Geoscience for Natural Gas Subsurface Systems* (pp. 345–382). Elsevier.
- Rahman, M. K., Suarez, Y. A., Chen, Z., & Rahman, S. S. (2007). Unsuccessful hydraulic fracturing cases in Australia: Investigation into causes of failures and their remedies. *Journal of Petroleum Science and Engineering*, 57(1–2), 70–81.
- Safari, M. R., Trevor, O., Queen, C., Hamed, C., Blair, N., Uno, M., & Hawkes, C. D. (2013). Effects of Depletion/Injection Induced Stress Changes on Natural Fracture Reactivation. *47th U.S. Rock Mechanics/Geomechanics Symposium*.
- Song, D., Wang, X., Wu, C., Meng, S., Zhang, M., Li, H., Jiao, H., Liu, X., Jin, X., & Tuo, J. (2021). Petroleum Generation, Retention, and Expulsion in Lacustrine Shales Using an Artificial Thermal Maturation Approach: Implications for the In-Situ Conversion of Shale Oil. *Energy & Fuels*, 35(1), 358–373.
- Srivardhan, V., & Pal, S. (2015). Shale Volume Estimation Using Factor Analysis and Neural Network. *6th International Geosciences Student Conference*.
- Szabó, N. P. (2011). Shale volume estimation based on the factor analysis of well-logging data. *Acta Geophysica*, 59(5), 935–953.
- Szabó, N. P., & Dobróka, M. (2018). Exploratory Factor Analysis of Wireline Logs Using a Float-Encoded Genetic Algorithm. *Mathematical Geosciences*, 50(3), 317–335.
- Tahmasebi, P., Javadpour, F., & Sahimi, M. (2017). Data mining and machine learning for identifying sweet spots in shale reservoirs. *Expert Systems with Applications*, 88, 435–447.

- Tenaglia, M., Eberli, G. P., Weger, R. J., Blanco, L. R., Rueda Sánchez, L. E., & Swart, P. K. (2020). Total organic carbon quantification from wireline logging techniques: A case study in the Vaca Muerta Formation, Argentina. *Journal of Petroleum Science and Engineering*, 194, 107489.
- Venieri, M., Harazim, D., Pedersen, P. K., & Eaton, D. W. (2021). Vertical and lateral facies variability in organic-rich mudstones at the reservoir scale: A case study from the Devonian Duvernay formation of Alberta, Canada. *Marine and Petroleum Geology*, 132, 105232.
- Vikara, D., Remson, D., & Khanna, V. (2020). Gaining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling. *Upstream Oil and Gas Technology*, 4, 100007.
- Wanniarachchi, W. A. M., Ranjith, P. G., Perera, M. S. A., Lashin, A., Al Arifi, N., & Li, J. C. (2015). Current opinions on foam-based hydro-fracturing in deep geological reservoirs. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 1(3–4), 121–134.
- Xie, J. (2018). Rapid shale gas development accelerated by the progress in key technologies: A case study of the Changning–Weiyuan National Shale Gas Demonstration Zone. *Natural Gas Industry B*, 5(4), 283–292.
- Yu, R. (2020). Application of big data in oil and gas exploration and development-shale gas field in south Sichuan as an example. *Mineral Exploration*, 9, 2000–2007.
- Zhang, P., Lu, S., & Li, J. (2019). Characterization of pore size distributions of shale oil reservoirs: A case study from Dongying sag, Bohai Bay basin, China. *Marine and Petroleum Geology*, 100, 297–308.
- Zhao, P., Mao, Z., Huang, Z., & Zhang, C. (2016). A new method for estimating total organic carbon content from well logs. *AAPG Bulletin*, 100(08), 1311–1327.
- Zhou, Z., Gunter, W. D., Kadatz, B., & Cameron, S. (1996). Effect Of Clay Swelling On Reservoir Quality. *Journal of Canadian Petroleum Technology*, 35(07).
- Zou, C., Xue, H., Xiong, B., Zhang, G., Pan, S., Jia, C., Wang, Y., Ma, F., Sun, Q., Guan, C., & Lin, M. (2021). Connotation, innovation and vision of “carbon neutrality”. *Natural Gas Industry B*, 8(5), 523–537.
- Zou, C., Yang, Z., Cui, J., Zhu, R., Hou, L., Tao, S., Yuan, X., Wu, S., Lin, S., Wang, L., Bai, B., & Yao, J. (2013). Formation mechanism, geological characteristics and development strategy of nonmarine shale oil in China. *Petroleum Exploration and Development*, 40(1), 15–27.

Acknowledgments

I am profoundly grateful for the support and guidance I have received throughout the journey of completing this OTDK thesis. This accomplishment would not have been possible without numerous individuals' unwavering encouragement and assistance.

First, I sincerely appreciate Professor Dr. Norbert Péter Szabó, my esteemed supervisor, for his invaluable expertise and unwavering support throughout this research endeavor. His insightful feedback, patience, and encouragement have been instrumental in shaping this thesis.

I want to also thank my mother institution, the University of Maiduguri, Borno State, Nigeria, for granting me a study leave to come this far for studies, the Vice Chancellor, Registrar, and all staff of Geology Department, University of Maiduguri, most especially Professor Zhao Peiqiang, Professor Saidu Baba, Dr. Mohammed Bukar, Mal. Bashiru and Late. Professor Shettima Bukar Ahmed, Mr. Rowland, and Mr. Edoson for allowing me to be in the academic system and the transfer of knowledge I gained from them.

I am deeply indebted to my beloved family, especially my late father; Engr. Sylvester Obidiegwu, my late mother, Mrs. Juliet Obidiegwu and my lovely wife, Mrs. Alice Martins Obidiegwu (Kulliss), whose boundless love, understanding, and encouragement have been a constant source of strength and motivation. Their sacrifices and unwavering belief in me have been the cornerstone of my success.

I extend heartfelt gratitude to my dear friends and brothers, Mr. Richards Ekene Obidiegwu, Mr. Zhang Yongjia, Dr. Kiyangi Wyclif, Dr. Azam Muhammed Khan, Dr. Mandella Ali Fargalla, Dr. Edoson Enesto, Abigail, Jean, Amaoku Kojo, Richmon, Mayowa Babayemi Temitope, Mrs. Temitope Mayowa Babayemi, Josephmartin, Mahmood, Azeez Sobowale, Abdulkaleed, Moroj, Akos Laszlo, Laura Kiraly, Anthonio and Benjamin Chomba, for their unwavering support, mentorship, encouragement, and camaraderie throughout this academic journey. Their friendship, wise counsel and encouragement have been invaluable in navigating the challenges of academia. I am also grateful to all my other friends and colleagues who have provided support, inspiration, and valuable insights during this research thesis.

Lastly, I would like to express my gratitude to the Scholarship for Christian Young People (SCYP), University of Miskolc, the TDK conference organizers, especially the secretary, Dr. Anna Boglarka, the Chinese Government Scholarship program, China University of Petroleum, Beijing, and all staff of the College of Geophysics for providing me with the opportunity to pursue my academic aspirations and for their generous support throughout my studies.

Each of you has significantly shaped this thesis and my academic journey, and I am truly grateful for that.