Algerian Democratic and Popular Republic
Ministry of Higher Education and Scientific Research

**KASDI MERBAH UNIVERSITY OUARGLA**

Faculty of Hydrocarbons, Renewable Energies, and Earth and Universe Sciences
Department of Earth and Universe Sciences

## Master Dissertation

**Specialization: PETROLEUM GEOLOGY**

# AI-Based Prediction of Rock Petrophysical Characteristics from Well Logs

Presented by:

**ISSAM BOUGHABA**
**MANSOUR BOUAZZA**
**FATIMA CHAOUCH**

Supervisors:
**Benkhelifa Randa & Nemer Zoubida**

Jury

| Name | | Institution | Role |
|---|---|---|---|
| ASMA CHEMAM | MCB | UKM Ouargla | **Jury Chair** |
| MOHAMED SALAH BELKSEIR | Professor | UKM Ouargla | **Examiner** |
| RANDA BENKHELIFA | MCB | UKM Ouargla | **Supervisor** |
| ZOUBIDA NEMER | MCB | UKM Ouargla | **Co-supervisor** |

**Academic Year: 2025 / 2026**

# ACKNOWLEDGMENTS

Before presenting this humble work, we would like to express our sincere gratitude and appreciation to all those who contributed directly or indirectly to its completion.

First and foremost, we thank Almighty God for granting us the strength and courage to pursue and complete our studies.

We would like to express our deepest appreciation to our supervisors, Dr. ZOUBIDA NEMER and Dr. RANDA BENKHELIFA, lecturers at Kasdi Merbah University of Ouargla, for their constant support, valuable guidance, and availability throughout the preparation of this dissertation.

We also extend our sincere thanks to the members of the jury, Dr. ASMA CHEMAM and Prof. BELKSEIR Mohamed Salah (Examiner), for the honor they have bestowed upon us by evaluating this work.

We would also like to thank our families and loved ones for their moral support and continuous encouragement, which have been an invaluable source of motivation.

Finally, our sincere thanks go to all our teachers, colleagues, and everyone who contributed, directly or indirectly, to the successful completion of this dissertation.

# ABSTRACT

This study evaluates and compares five supervised machine learning regression algorithms , Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR), and Extreme Gradient Boosting (XGBoost), for the prediction of three key petrophysical parameters (effective porosity, water saturation, and shale volume) from wireline well log data in the Triassic Argilo-Gréseux Inférieur (TAGI) reservoir of the Berkine Basin, eastern Algeria. The study uses a dataset of 12 wells with 32 measured and interpreted log parameters. XGBoost consistently outperforms competing algorithms, achieving $R^2$ values of 0.93–0.96 through optimized 5-fold cross-validation. SHAP-based feature importance analysis confirms the physical consistency of model predictions, identifying gamma ray, resistivity suite, and sonic transit time as the dominant predictors. The methodology provides a validated, reproducible workflow for data-driven reservoir characterization applicable to analogous Triassic sandstone systems in North Africa.

**Keywords:** *machine learning; XGBoost; petrophysical prediction; TAGI reservoir; Berkine Basin; well logs; SHAP interpretability*

# RÉSUMÉ

Cette étude évalue et compare cinq algorithmes de régression par apprentissage automatique supervisé, Régression Linéaire Multiple (MLR), Arbre de Décision (DTR), Forêt Aléatoire (RFR), Gradient Boosting (GBR) et XGBoost, pour la prédiction de trois paramètres pétrophysiques clés (porosité effective, saturation en eau et volume d'argile) à partir de données de diagraphie dans le réservoir Triasique Argilo-Gréseux Inférieur (TAGI) du Bassin de Berkine, Algérie orientale. L'étude utilise un jeu de données de 12 puits avec 32 paramètres mesurés et interprétés. XGBoost surpasse systématiquement les algorithmes concurrents, atteignant des valeurs de $R^2$ de 0,93 à 0,96. L'analyse d'importance des variables par SHAP confirme la cohérence physique des prédictions du modèle.

**Mots-clés :** *apprentissage automatique ; XGBoost ; prédiction pétrophysique ; réservoir TAGI ; Bassin de Berkine ; diagraphies de puits ; interprétabilité SHAP*

# TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... 2
ABSTRACT ..... 3
RÉSUMÉ ..... 4
TABLE OF CONTENTS ..... 5
LIST OF FIGURES ..... 8
LIST OF TABLES ..... 9
LIST OF ABBREVIATIONS ..... 10
GENERAL INTRODUCTION ..... 12
Context and Problem Statement ..... 12
Research Objectives ..... 12
Thesis Structure ..... 13
CHAPTER 1: GEOLOGICAL SETTING ..... 14
1.1 Geological Setting of the Berkine Basin ..... 14
1.1.1 Regional Location and Basin Geometry ..... 14
1.1.2 Tectonic Evolution and Structural Framework ..... 15
1.2 Stratigraphy of the Berkine Basin ..... 16
1.2.1 General Stratigraphic Framework ..... 16
1.2.2 The TAGI Reservoir (Depositional Setting and Facies) ..... 17
1.2.3 Diagenesis and Reservoir Quality ..... 18
1.3 Petroleum Systems of the Berkine Basin ..... 19
1.3.1 Source Rocks ..... 19
1.3.2 Migration, Traps, and Seals ..... 20
1.3.3 Low-Resistivity Pay Zones ..... 20
1.4 Petrophysical Properties of the TAGI Reservoir ..... 20
CHAPTER 2: THEORETICAL BACKGROUND ..... 22
2.1 Fundamentals of Petrophysical Parameters ..... 22
2.1.1 Porosity ..... 22
2.1.2 Water Saturation ..... 22
2.1.3 Shale Volume ($V_{CL}$) ..... 23
2.1.4 Permeability and Indirect Estimation ..... 23
2.2 Wireline Well Logging Principles ..... 24
2.2.1 Fundamentals of Log Acquisition ..... 24
2.2.2 Log Suite Summary ..... 24
2.2.3 Gamma Ray and Spectral Gamma Ray Logs ..... 25
2.2.4 Density, Neutron, and Acoustic Logs ..... 25

2.3 Statistical and Machine Learning Concepts ..... 26
2.3.1 Supervised Learning and Regression ..... 26
2.3.2 Data Splitting and Cross-Validation ..... 26
2.3.3 Overfitting Diagnosis ..... 26
2.4 Regression Algorithms ..... 26
2.4.1 Multiple Linear Regression (MLR) ..... 27
2.4.2 Decision Tree Regression (DTR) ..... 27
2.4.3 Bagging Regressor ..... 28
2.5 Ensemble Learning Methods ..... 28
2.5.1 Random Forest Regression (RFR) ..... 28
2.5.2 Extremely Randomised Trees (ExtraTree) ..... 28
2.5.3 Gradient Boosting Regression (GBR) ..... 28
2.5.4 Extreme Gradient Boosting (XGBoost) ..... 29
2.6 Model Evaluation Metrics ..... 29
2.7.1 Neural Network Approaches ..... 29
2.7.2 Ensemble Tree-Based Methods ..... 30
2.7.3 Deep Learning and Berkine Basin Benchmarks ..... 30
CHAPTER 3: METHODOLOGY ..... 31
3.1 Dataset Description and Well Inventory ..... 31
3.1.1 Study Wells ..... 31
3.1.2 Raw Data Structure ..... 32
3.2 Data Preprocessing and Quality Control ..... 32
3.2.1 Null Value Treatment ..... 32
3.2.2 Depth Alignment and Interval Selection ..... 32
3.2.3 Outlier Detection and Treatment ..... 33
3.2.4 Log-Transformation of Resistivity ..... 33
3.2.5 Median Imputation of Missing Feature Values ..... 33
3.2.6 Data Validity Filter ..... 33
3.3 Feature Selection ..... 34
3.3.1 NaN Threshold Criterion ..... 34
3.3.2 Retained Feature Set and Physical Justification ..... 35
3.4 Train/Test Split Strategy ..... 35
3.4.1 Well-Based Split Rationale ..... 35
3.4.2 Split Criteria and Well Assignment ..... 36
3.5 Model Development and Implementation ..... 37
3.5.1 Algorithm Selection ..... 37
3.5.2 Visualization Outputs ..... 38

CHAPTER 4: RESULTS AND DISCUSSION ..... 39
4.1 Exploratory Data Analysis ..... 39
4.1.1 Spearman Correlation Matrix ..... 39
4.1.2 Feature Importance ..... 40
4.2 Water Saturation ($S_W$) Prediction ..... 41
4.2.1 Comparative Model Performance ..... 41
4.2.2 Predicted vs. Actual Distributions ..... 42
4.2.3 Cross-diagram Analysis ..... 43
4.2.4 Per-well Analysis and the Well 9 Anomaly ..... 43
4.3 Clay Volume ($V_{CL}$) Prediction ..... 44
4.3.1 Comparative Model Performance ..... 44
4.3.2 Distributions and Cross-diagram ..... 45
4.3.3 Per-well Analysis ..... 45
4.4 Effective Porosity (PORO) Prediction ..... 46
4.4.1 Comparative Model Performance ..... 46
4.4.2 Cross-diagram and Distribution Analysis ..... 47
4.4.3 Per-well Analysis ..... 47
4.5 Depth Profiles and Residual Analysis ..... 48
4.6 Spatial Synthesis and Well Performance ..... 51
4.6.1 Box Plot Comparison ..... 51
4.6.2 $R^2$ Performance Map by Well and Target ..... 52
4.7 General Discussion ..... 52
CONCLUSION AND PERSPECTIVES ..... 54
REFERENCES ..... 57

# LIST OF FIGURES

Figure 1: Location map of the Ourhoud field in the Berkine Basin, Algeria. (Rossi et al.; 2002). ..... 15
Figure 2: Generalized stratigraphic column of the Berkine Basin (Rossi et al., 2002). ..... 17
Figure 3: NNW–SSW generalized cross-section through the Berkine Basin (Boote et al. (1998) in Rossi et al., 2002)..... 19
Figure 4: Spearman correlation matrix for SW training wells. ..... 40
Figure 5: Feature importance for SW prediction (Random Forest). ..... 41
Figure 6: Actual vs. predicted distributions for SW, VCL, and PORO. ..... 42
Figure 7: Cross-diagram XGBoost SW predicted vs. actual. ..... 43
Figure 8: Cross-diagram XGBoost VCL predicted vs. actual. ..... 45
Figure 9: Cross-diagram XGBoost PORO predicted vs. actual. ..... 47
Figure 10: Seven-track depth profile: Well 10. Bias: ΔSW = +0.040, ΔVCL = +0.007, ΔPORO ≈ 0.000. ..... 48
Figure 11: Seven-track depth profile: Well 11. Bias: ΔSW = +0.024, ΔVCL = +0.056, ΔPORO = −0.012. ..... 49
Figure 12: Seven-track depth profile: Well 12. Bias: ΔSW = +0.054, ΔVCL = +0.061, ΔPORO = +0.004. ..... 50
Figure 13: Seven-track depth profile: Well 9. Severe ΔSW bias (+0.283) confirms SW = 1.0 anomaly. VCL acceptable (bias = −0.025). ..... 51
Figure 14: Box plots of actual vs. XGBoost predicted SW and VCL per test well. ..... 52
Figure 15: XGBoost $R^2$ per test well for SW, VCL, and PORO. ..... 52

# LIST OF TABLES

Table 1: Summary of TAGI reservoir petrophysical properties compiled from published literature. Pay zone values refer to intervals meeting minimum cutoffs for net pay classification (φ ≥ 10%, Sw ≤ 60%, VCL ≤ 35%). ..... 21
Table 2: Well log suite with physical measurement principles and petrophysical applications. .. 24
Table 3: Comparative summary of the seven regression algorithms: type, core mechanism, strengths, and limitations in the context of petrophysical log prediction. ..... 27
Table 4: Well inventory of the 12 study wells with field location, TAGI depth range, log suite availability, and train/test assignment. Blue rows = blind test wells. ..... 31
Table 5: Feature selection results for all 15 candidate log inputs. NaN percentage computed across all 12 wells within TAGI reservoir intervals. Green = retained; Red = excluded. ..... 34
Table 6: Final well-based train/test split. ..... 36
Table 7: Seven regression algorithms compared in this study with key hyperparameter configurations, type, strengths, and limitations in the context of petrophysical log prediction. ... 37
Table 8: Performance comparison of seven regression algorithms for SW prediction on blind-test wells (n = 21,678). ..... 42
Table 9: XGBoost SW performance by test well. ..... 44
Table 10: Performance comparison of seven regression algorithms for VCL prediction (n = 22,293)..... 44
Table 11: XGBoost VCL performance by test well. ..... 46
Table 12: Performance comparison of seven regression algorithms for PORO prediction (n = 22,364)..... 46
Table 13: XGBoost PORO performance by test well. ..... 48

# LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| **AI** | Artificial Intelligence |
| **ANN** | Artificial Neural Network |
| **AAPG** | American Association of Petroleum Geologists |
| **AARE** | Average Absolute Relative Error |
| **BVW** | Bulk Volume Water (φ × Sw) |
| **CALI** | Caliper Log |
| **CNN** | Convolutional Neural Network |
| **DTC** | Compressional Sonic Transit Time (μs/ft) |
| **DTS** | Shear Sonic Transit Time (μs/ft) |
| **DTR** | Decision Tree Regression |
| **GBR** | Gradient Boosting Regression |
| **GR** | Gamma Ray Log (API) |
| **GRTO** | Total Gamma Ray |
| **IGR** | Gamma Ray Index (shale indicator) |
| **LAS** | Log ASCII Standard |
| **MAE** | Mean Absolute Error |
| **ML** | Machine Learning |
| **MLR** | Multiple Linear Regression |
| **MSE** | Mean Square Error |
| **NPHI** | Neutron Porosity Log (fraction) |
| **OOB** | Out-of-Bag (Random Forest) |
| **PE** | Photoelectric Factor Log (barns/electron) |
| **PERM** | Permeability (mD) |
| **PHIE** | Effective Porosity (fraction) |
| **POR** | Total Porosity (fraction) |
| **POTA** | Potassium Spectral Log (%) |
| **$R^2$** | Coefficient of Determination |
| **RF** | Random Forest |
| **RFR** | Random Forest Regression |
| **RHOB** | Bulk Density Log (g/cm³) |
| **RMSE** | Root Mean Square Error |
| **RT** | True Formation Resistivity (ohm·m) |

| | |
|---|---|
| **SHAP** | SHapley Additive exPlanations |
| **SONATRACH** | Société Nationale pour la Recherche, la Production, le Transport, la Transformation et la Commercialisation des Hydrocarbures |
| **SW** | Water Saturation (fraction) |
| **TAGI** | Triassic Argilo-Gréseux Inférieur (main Berkine Basin reservoir) |
| **THOR** | Thorium Spectral Log (ppm) |
| **TVD** | True Vertical Depth (m) |
| **URAN** | Uranium Spectral Log (ppm) |
| **VCL** | Volume of Clay / Shale Volume (fraction) |
| **XGBoost** | Extreme Gradient Boosting |

# GENERAL INTRODUCTION

## Context and Problem Statement

Reservoir characterization is the cornerstone of petroleum engineering, providing the quantitative description of subsurface rock properties that govern fluid storage and flow. Among all reservoir parameters, petrophysical properties, porosity (φ), permeability (k), water saturation (Sw), and shale volume ($V_{cl}$), are the most critical determinants of hydrocarbon volumes in place and producibility. Conventionally, these properties are derived from wireline well log data using empirical equations such as Archie's equation for water saturation, density–neutron crossplots for porosity, and gamma ray linear transforms for shale volume (Archie, 1942; Schlumberger, 1989).

The Berkine Basin (Ghadames Basin) in eastern Algeria represents one of North Africa's most productive petroleum provinces. Since the major discoveries of the 1990s, including the Hassi Berkine, El Merk, and Ourhoud fields, the basin has contributed a significant fraction of Algeria's oil and gas reserves. The principal producing reservoir is the Triassic Argilo-Gréseux Inférieur (TAGI), a continental fluvial sandstone sequence deposited unconformably above the Hercynian erosion surface (Rossi et al., 2002; Galeazzi et al., 2009). Despite decades of exploration and production, detailed petrophysical characterization of TAGI reservoir intervals remains challenging due to lithological heterogeneity, variable clay content, and diagenetic overprints that complicate the application of standard log interpretation models.

A critical limitation in the dataset compiled for this study is the complete absence of measured permeability values (PERM = −999.25, the standard LAS null flag) across all 12 wells. The present study therefore focuses on direct AI prediction of porosity, water saturation, and shale volume, all available as conventionally interpreted curves in the dataset, and uses the predicted values as inputs to a secondary permeability proxy model (Miah et al., 2024).

## Research Objectives

The primary objectives of this thesis are:

- To evaluate and compare five AI regression algorithms (MLR, DTR, RFR, GBR, XGBoost) for predicting porosity, water saturation, and shale volume from wireline well log data in the Berkine Basin TAGI reservoir.
- To identify the most influential well log inputs for each petrophysical target through SHAP-based feature importance analysis, ensuring physical interpretability of the AI models.
- To develop a validated, reproducible machine learning workflow applicable to reservoir characterization in the Berkine Basin and analogous Triassic sandstone reservoirs in North Africa.
- To provide an indirect permeability estimate using the Timur-Coates empirical equation calibrated with AI-predicted porosity and irreducible water saturation.
- To benchmark results against published AI petrophysical studies from comparable basins worldwide.

## Thesis Structure

This thesis is organized into four main chapters preceded by a general introduction and followed by a conclusion with perspectives for future research. Chapter 1 reviews the geological and stratigraphic context of the Berkine Basin and the state of the art in AI-based petrophysical prediction. Chapter 2 provides the theoretical background on petrophysical parameters, wireline logging, and machine learning regression concepts. Chapter 3 describes the complete methodology from data acquisition and preprocessing to model development, cross-validation, and SHAP analysis. Chapter 4 presents and discusses the results, including model comparisons, feature importance findings, and their geological interpretation in the context of the TAGI reservoir.

# CHAPTER 1: GEOLOGICAL SETTING

## 1.1 Geological Setting of the Berkine Basin

### 1.1.1 Regional Location and Basin Geometry

The Berkine Basin, also referred to as the eastern sector of the Ghadames Basin, is an intracratonic sedimentary basin situated in the Saharan Platform of eastern Algeria, extending into northwestern Libya and southern Tunisia (Fig.1). It covers approximately 250,000 km² in its Algerian sector alone, with a maximum sedimentary fill of up to 7,000 m in the depocenters (Galeazzi et al., 2009). The basin is elongated roughly in the NNW–SSE direction and is bounded to the west by the north–south trending Hassi Messaoud Ridge, which separates it from the Oued Mya Basin, and to the east by the Libyan border faults (Galeazzi et al., 2009). The Amguid–Hassi Messaoud Arch forms a major structural axis along the western margin, while to the north the Berkine High divides the basin from the Illizi Basin.

The Berkine Basin sits at the convergence of several major structural lineaments of the Saharan Platform, making it one of the most tectonically complex intracratonic settings in North Africa. The basin's depocenters record more than 500 million years of sedimentation, with the deepest part of the Paleozoic succession reaching approximately 5,500 m below sea level in the eastern sub-basins (Galeazzi et al., 2009). This exceptional sedimentary thickness, combined with a favorable burial history, created ideal conditions for the generation, migration, and trapping of large hydrocarbon volumes. Since major commercial discoveries were made in the 1990s, including the Hassi Berkine, El Merk, and Ourhoud fields, the basin has consistently contributed more than 30% of Algeria's total hydrocarbon production (Moffat et al., 2001).

.

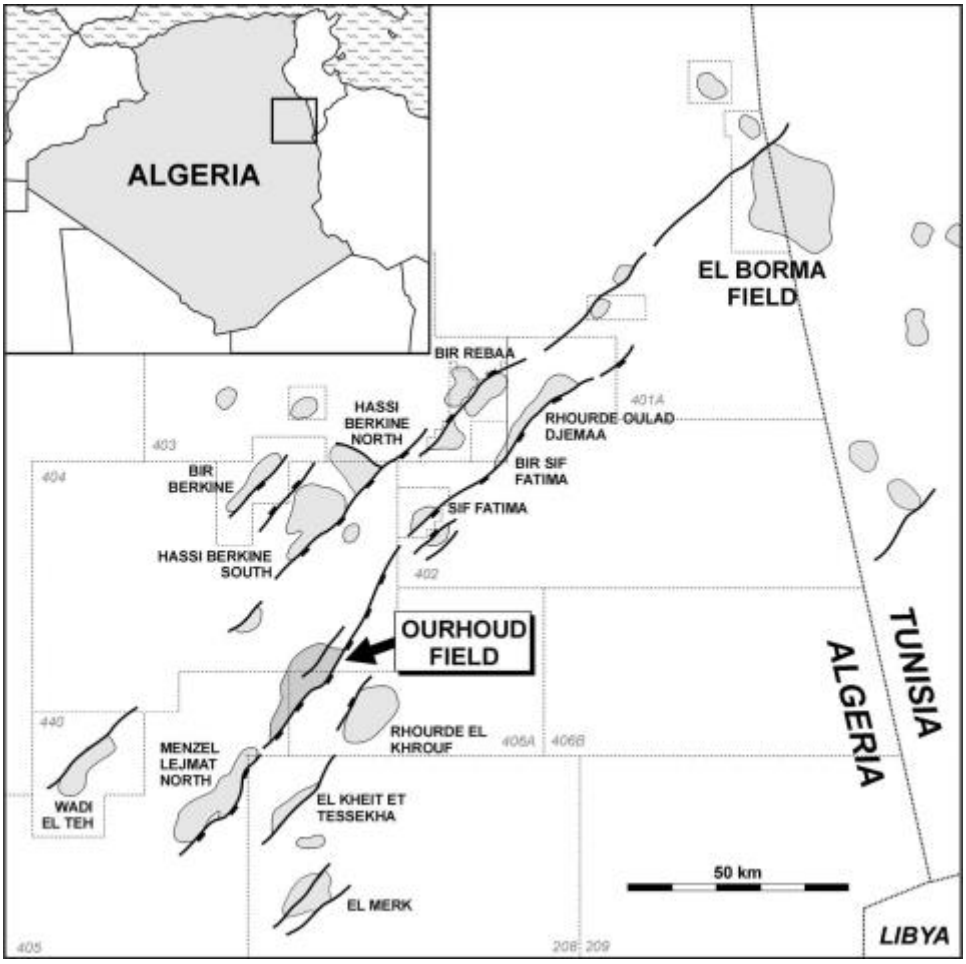


*Figure 1:Location map of the Ourhoud field in the Berkine Basin, Algeria. (Rossi et al.; 2002).*

### 1.1.2 Tectonic Evolution and Structural Framework

The structural architecture of the Berkine Basin was shaped by at least four major tectonic episodes spanning the Proterozoic to the Cenozoic. Understanding these events is important for interpreting the present-day reservoir geometry, trap configurations, and fluid migration pathways.

The Pan-African Orogeny (700–550 Ma) assembled the basement of the Saharan Craton through collision and suturing of terranes. The resulting basement fabric, composed of Precambrian metamorphic and igneous rocks, exerts a first-order control on the orientation of Phanerozoic faults and fracture systems. Reactivation of Pan-African basement lineaments during subsequent tectonic episodes produced the predominant NE–SW and NW–SE structural grain visible in seismic attribute maps across the basin (Boote et al., 1998).

The Caledonian phase in the Early Paleozoic produced moderate compressional deformation in the northern Sahara but had limited impact on the Berkine area. More significant was the Hercynian Orogeny, which culminated in the late Carboniferous to early Permian (approximately 300–280 Ma). This event produced the most important regional unconformity in North Africa, the Hercynian erosional surface, which truncated Paleozoic rocks at varying depths across the Algerian Sahara. In the Berkine Basin, the Hercynian erosion removed the Carboniferous section and part of the Devonian over structural highs, but left a thick Paleozoic succession in the basin depocenters. This

unconformity surface directly controlled the distribution and thickness of the overlying Triassic continental sediments (Turner et al., 2001).

Triassic–Jurassic rifting associated with the early opening of the Tethys Ocean created a series of NE–SW trending extensional faults and half-grabens that define the main structural closures in the TAGI reservoir. These fault-bounded blocks constitute the principal trap style in the Berkine Basin, supplemented by subtle four-way dip closures on regional anticlines. The Cretaceous period brought mild regional compression that tightened pre-existing structural closures without creating major new structural elements. Final Alpine compression in the Cenozoic tilted the entire Saharan Platform slightly northward and reactivated some pre-existing reverse faults along the basin margins (Galeazzi et al., 2009).

## 1.2 Stratigraphy of the Berkine Basin

### 1.2.1 General Stratigraphic Framework

The stratigraphic succession of the Berkine Basin encompasses rocks from the Cambro-Ordovician to the Quaternary and represents a nearly complete record of Phanerozoic sedimentation on the North African Saharan Platform (Galeazzi et al., 2009; Turner et al., 2001). Table 1.1 summarizes the principal stratigraphic units and their petroleum system roles. The succession can be divided into two megasequences separated by the Hercynian unconformity: the Gondwana megasequence (Cambrian to Carboniferous) deposited during the long Paleozoic passive margin phase, and the Tethys megasequence (Triassic to Cretaceous) deposited following continental rifting and the gradual northward drift of the African plate.

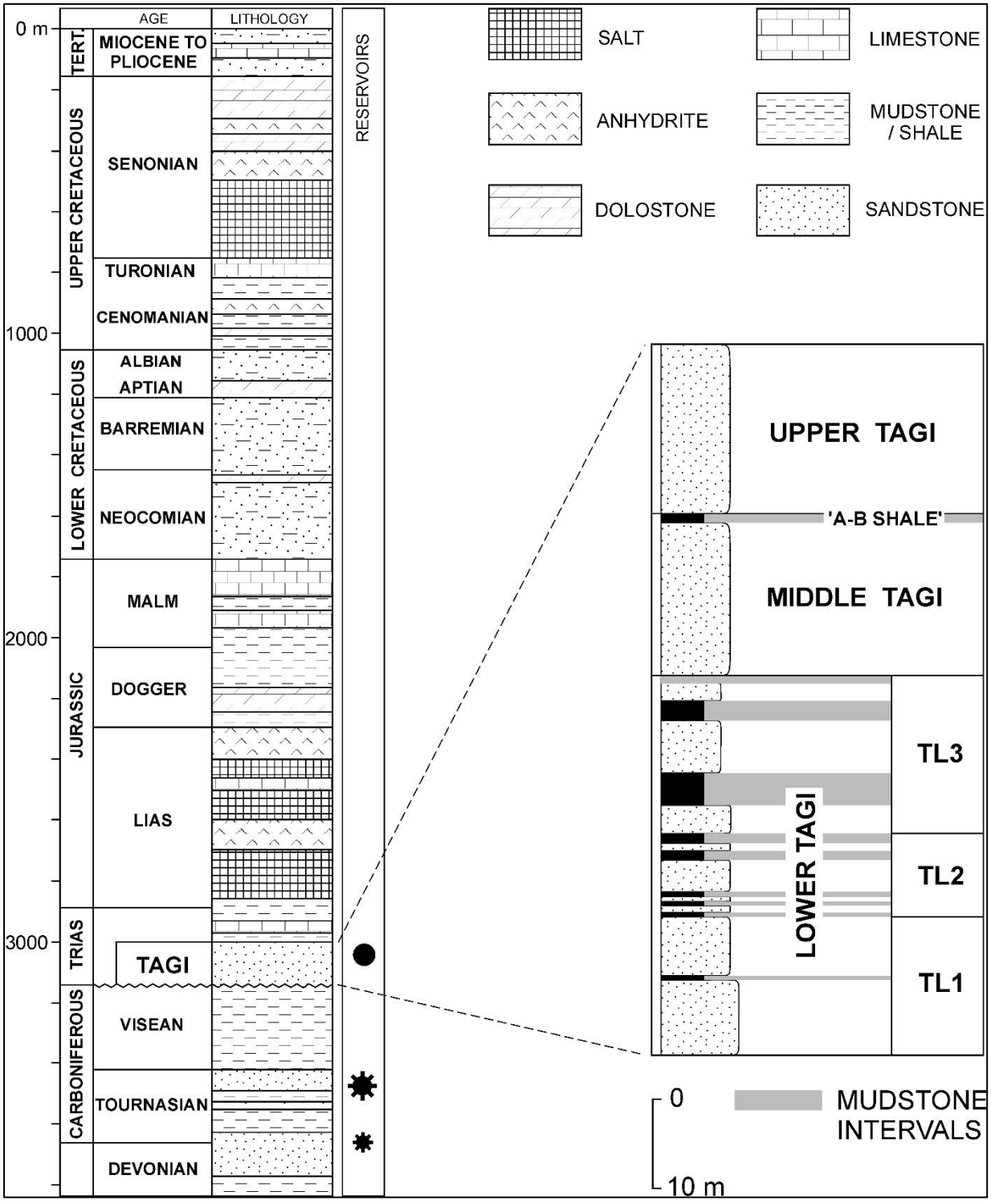


*Figure 2:Generalized stratigraphic column of the Berkine Basin (Rossi et al, 2002)*

### 1.2.2 The TAGI Reservoir (Depositional Setting and Facies)

The Triassic Argilo-Gréseux Inférieur (TAGI) is the principal producing reservoir in the Berkine Basin and the focus of the present study. It consists of a continental fluvial–alluvial sandstone sequence deposited unconformably above the Hercynian erosion surface during the Early to Middle Triassic (approximately 247–235 Ma). The TAGI was deposited in a broad, low-relief

continental interior basin with a semi-arid to arid climate, fed by paleodrainage systems flowing from Precambrian basement highs to the south and east (Turner et al., 2001).

The depositional architecture of the TAGI includes three main facies associations. The first is braided fluvial channel facies, which consists of coarse-to-medium-grained, well-sorted quartzarenites with tabular and trough cross-bedding, forming the best-quality reservoir units with porosities of 18–25% and permeabilities exceeding 200 mD in the least-cemented intervals. The second facies association is floodplain and overbank deposits, consisting of fine-grained sandstones and red-brown siltstones and claystones that separate the channel sands and constitute the intra-reservoir shale baffles. The third association is alluvial fan and sheet-flood deposits, predominantly present near basement paleohighs where coarser, poorly sorted conglomeratic sediments form localized, lower-quality reservoir units (Rossi et al., 2002).

The net-to-gross ratio (N/G) of the TAGI varies significantly across the basin, ranging from 30% in the more distal, floodplain-dominated areas to 80% in the proximal, channel-dominated zones. This lateral heterogeneity is a major driver of reservoir quality variation and is the primary reason that well-by-well petrophysical interpretation, rather than basin-wide empirical averages, is required for accurate resource estimation (Turner et al., 2001).

### 1.2.3 Diagenesis and Reservoir Quality

Diagenetic processes profoundly modify the primary depositional porosity and permeability of the TAGI sandstones and create the heterogeneous reservoir quality distribution that motivates the AI-based approach of this study. Rossi et al. (2002) conducted a comprehensive petrographic and geochemical study of TAGI samples from the Ourhoud field, documenting six principal diagenetic processes operating in approximate paragenetic sequence:

- Mechanical compaction during burial to 500–1,500 m: rearrangement and deformation of grain contacts reduce primary porosity from approximately 35–40% (depositional) to 25–30%.
- Quartz cementation (syntaxial overgrowths): the dominant porosity-reducing process, reducing porosity by 5–12% in deeply buried (>2,500 m) samples. Quartz cementation is strongly temperature-controlled and accelerates above 80°C, creating a depth-dependent porosity reduction gradient.
- Illitic clay coatings on detrital quartz grains: early diagenetic coatings partially inhibit quartz overgrowth development by preventing epitaxial nucleation, preserving anomalously high porosity in some deeply buried samples. Clay coating is most developed in the coarser-grained channel facies.
- Kaolinite-to-dickite transformation: kaolinite precipitated by meteoric water flushing during the Hercynian uplift transforms to dickite upon burial reheating above 100°C. Both phases partially occlude pore throats, reducing permeability disproportionately relative to porosity.
- Fe-dolomite cementation: localized dolomite cement patches associated with Mg-rich formation waters from evaporitic source beds completely occlude pores in affected zones, creating tight 'super-K' baffles within otherwise porous sands.

- Feldspar dissolution: dissolution of detrital K-feldspars creates secondary micro-porosity that partially offsets the porosity lost to cementation but contributes little to effective permeability due to the small size and poor connectivity of dissolution pores.

The combined effect of these diagenetic processes is a reservoir with significant vertical and lateral heterogeneity in petrophysical properties, where adjacent samples from the same well can show porosity variations of 10–15% and permeability variations of two to three orders of magnitude. This heterogeneity exceeds the capacity of deterministic empirical log interpretation methods to resolve accurately.

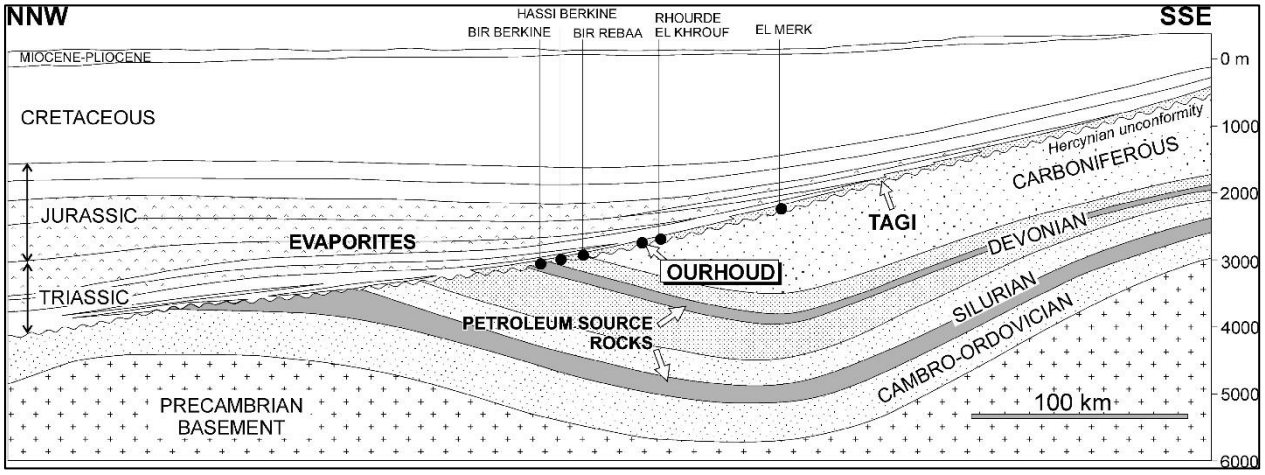


*Figure 3: NNW–SSW generalized cross-section through the Berkine Basin (Boote et al. (1998) in Rossi et al., 2002).*

## 1.3 Petroleum Systems of the Berkine Basin

### 1.3.1 Source Rocks

Yahi et al. (2001) conducted a comprehensive geochemical and basin modeling study identifying two primary source rock intervals that charge the TAGI reservoir. The primary source is the Llandoverian–Wenlockian Silurian Tannezuft Formation (also known as the "hot shales" due to their elevated uranium content and associated high gamma-ray log response). This unit consists of organic-rich black shales with total organic carbon (TOC) contents of 2–4%, deposited in an anoxic distal shelf to basin setting during the maximum Silurian marine transgression. The Tannezuft is the dominant source rock for the vast majority of Algerian Saharan oil fields and is estimated to have generated 80–90% of the hydrocarbons in the Berkine Basin. Vitrinite reflectance values of 0.8–1.2% in the basin center indicate that the Tannezuft is within the oil generation window ("oil kitchen")

The secondary source is the Frasnian Devonian black shales (F4 Formation), which have TOC contents of 1–2% and reflect deposition in restricted anoxic shelf environments during the Late Devonian Kellwasser oceanic anoxic events. The Devonian source contributes a secondary mixed oil signature to fields in the central and eastern parts of the basin. Basin modeling indicates Late Cretaceous to Tertiary timing for the main phase of oil generation and expulsion from both source intervals, consistent with the burial and thermal history controlled by the Mesozoic sedimentary loading

### 1.3.2 Migration, Traps, and Seals

Hydrocarbon migration from the Silurian and Devonian source kitchens to the TAGI reservoir occurred primarily through vertical fault pathways and short-distance lateral migration along the Hercynian unconformity surface. The unconformity functions as a migration highway because the Triassic continental sediments thin dramatically over Paleozoic structural highs, bringing the TAGI into near contact with the Devonian shales and reducing the vertical migration distance (Galeazzi et al., 2009). The trap inventory of the Berkine Basin is dominated by structural closures: fault-bounded horst blocks, rollover anticlines above extensional faults, and subtle four-way dip closures on broad regional anticlines. Stratigraphic traps associated with TAGI channel sand pinch-outs against the Hercynian unconformity are locally important but less well documented (Moffat et al., 2001).

The regional seal for the TAGI is provided by the overlying Triassic S1 evaporite sequence (halite and anhydrite), which forms a thick, laterally continuous pressure seal across most of the basin. The S1 seal is considered the primary factor explaining the exceptional preservation of large hydrocarbon accumulations in the Berkine Basin, as its low permeability and ductile mechanical behavior effectively prevent upward hydrocarbon leakage (Turner et al., 2007). Local intra-TAGI seals are provided by the floodplain shale and siltstone beds, which create internal baffles that subdivide the reservoir into multiple hydraulically distinct flow units relevant to production behavior and petrophysical zonation.

### 1.3.3 Low-Resistivity Pay Zones

Doghmane et al. (2024) conducted a dedicated study of the low-resistivity pay (LRP) problem in Algerian Saharan fields, a phenomenon particularly acute in the TAGI reservoir, demonstrating that electrically conductive clay minerals systematically reduce the resistivity of hydrocarbon-bearing intervals to values similar to or lower than adjacent water-saturated zones. This behavior, which violates the fundamental assumption of the Archie equation (that only water contributes to electrical conduction), leads to systematic underestimation of hydrocarbon saturation and, consequently, to misclassification of pay zones as non-productive intervals.

The LRP effect in the TAGI is primarily driven by three mechanisms: (1) dispersed illite and smectite clays within pore spaces that contribute excess conductivity through their high cation exchange capacity; (2) thin laminated shales at sub-log resolution that average resistivity downward in the tool response volume; and (3) high-salinity formation water (100–200 g/L NaCl equivalent) that depresses resistivity even in intervals with moderate-to-good hydrocarbon saturation. Doghmane et al. (2024) showed that artificial neural network (ANN) models trained on available log data can identify and correctly evaluate LRP intervals that conventional Archie-based models miss, recovering 15–25% additional net pay in tested wells.

## 1.4 Petrophysical Properties of the TAGI Reservoir

A comprehensive compilation of reported petrophysical properties from published studies of the TAGI reservoir provides the reference framework against which the AI model predictions of this study are benchmarked. Table 1 summarizes the key petrophysical parameters and their reported ranges.

*Table 1: Summary of TAGI reservoir petrophysical properties compiled from published literature. Pay zone values refer to intervals meeting minimum cutoffs for net pay classification ($\varphi \geq 10\%$, $Sw \leq 60\%$, $VCL \leq 35\%$).*

| Property | Range | Typical Pay Zone Value | Reference |
|---|---|---|---|
| **Total porosity ($\varphi_t$)** | 5–25% | 14–18% in pay zones | *Rossi et al. (2002)* |
| **Effective porosity ($\varphi_e$)** | 3–22% | 12–16% in pay zones | |
| **Permeability (k)** | 0.1–500 mD | 50–150 mD in clean sands | |
| **Water saturation (Sw)** | 15–100% | 25–55% in pay zones | *Doghmane et al. (2024)* |
| **Net-to-gross ratio** | 30–80% | Varies by depositional sub-environment | *Turner et al. (2001)* |
| **Reservoir temperature** | 80–120°C | Increases with depth (28–32°C/km) | *Yahi et al. (2001)* |
| **Reservoir pressure** | 250–380 bar | Normal to slightly overpressured | *Galeazzi et al. (2009)* |
| **API gravity (crude oil)** | 35–45° API | Light oil; low GOR | *Moffat et al. (2001)* |
| **Formation water salinity** | 100–200 g/L NaCl equiv. | High salinity; low Rw | *Doghmane et al. (2024)* |

The wide ranges reported in Table 1 reflect the diagenetic heterogeneity described in Section 1.2.3 and underscore the necessity of well-by-well, depth-specific petrophysical analysis. The relatively high formation water salinity (100–200 g/L NaCl equivalent) produces low Rw values (0.02–0.05 ohm·m at reservoir temperature), which partially explains the LRP behavior: when Rw is very low, even modest hydrocarbon saturation does not raise Rt sufficiently above the background water-wet resistivity to be unambiguously identified by conventional Archie interpretation (Doghmane et al., 2024).

This chapter has established the geological and stratigraphic framework within which the AI-based petrophysical prediction study is conducted. The TAGI reservoir presents a challenging petrophysical evaluation environment due to diagenetic heterogeneity, wide property ranges and the low-resistivity pay problem. The theoretical foundations of the petrophysical parameters and machine learning algorithms employed in this study are developed in Chapter 2.

# CHAPTER 2: THEORETICAL BACKGROUND

This chapter presents the theoretical foundations underpinning the methodology of this study. It is organised into eight sections: (i) the fundamental petrophysical parameters targeted for prediction; (ii) wireline well logging principles; (iii) statistical and machine learning concepts; (iv) the seven regression algorithms compared; (v) model evaluation metrics; (vi–viii) a structured review of published AI applications in petrophysical prediction.

## 2.1 Fundamentals of Petrophysical Parameters

### 2.1.1 Porosity

Porosity (φ) is defined as the ratio of void space volume to total bulk rock volume. Total porosity ($\varphi_t$) encompasses all pore spaces including isolated vugs and micro-pores within clay minerals, whereas effective porosity ($\varphi_e$) accounts only for interconnected pores that contribute to fluid storage and flow. In TAGI sandstones of the Berkine Basin, porosity ranges from below 5% in tightly cemented intervals to over 25% in well-sorted channel sands, with average values of approximately 14–18% in productive intervals (Rossi et al., 2002). Three independent log measurements are used to estimate porosity. The density-derived porosity (Compton scattering of gamma rays from a chemical source) is expressed as:

$$\varphi_D = (\rho_{ma} - \rho_b)/(\rho_{ma} - \rho_{fl})$$

where $\rho_{ma}$ is the matrix grain density (2.65 g/cm³ for quartz-dominated TAGI sands), $\rho_b$ is the measured bulk density from the RHOB log, and $\rho_{fl}$ is the pore fluid density (≈1.0 g/cm³ for brine, ≈0.8 g/cm³ for oil). The neutron-derived porosity (NPHI) responds to the hydrogen content of the formation: hydrogen atoms in pore fluids and clay-bound water slow and capture neutrons emitted from the tool source. The compressional sonic-derived porosity uses the Wyllie (1956) time-average equation:

$$\varphi_S = (DTC - DTC_{ma})/(DTC_{fl} - DTC_{ma})$$

where $DTC_{ma}$ is the matrix transit time (55 µs/ft for quartz) and $DTC_{fl}$ is the fluid transit time (189 µs/ft for water). In the present dataset, RHOB and NPHI are too sparse for model training, leaving DTC as the sole direct porosity indicator, a constraint that limits prediction accuracy and is discussed in Chapters 3 and 4.

### 2.1.2 Water Saturation

Water saturation (Sw) is the fraction of effective pore volume occupied by formation water; accordingly, hydrocarbon saturation is (1 − Sw). The classical model for clean, water-wet sandstones is Archie's equation (Archie, 1942):

$$Sw^n = (a \times Rw)/(\varphi^m \times Rt)$$

where Rw is the formation water resistivity (ohm·m), Rt is the true formation resistivity (deep resistivity RT90 in this dataset), a is the tortuosity factor (≈1 for consolidated sandstones), m is the cementation exponent (≈1.8–2.2 for sandstones), and n is the saturation exponent (≈2). In shaly sands, electrically conductive clay minerals violate the fundamental assumption that only pore

water conducts electrical current. Modified models (Waxman-Smits, Simandoux, Indonesia, dual-water) partially correct for this additional conductivity but may still be insufficient for low-resistivity pays, as demonstrated by Doghmane et al. (2024) in Algerian Saharan fields. This systematic limitation of deterministic interpretation in the TAGI directly motivates the AI-based approach of this study.

### 2.1.3 Shale Volume ($V_{CL}$)

Shale volume ($V_{CL}$) quantifies the proportion of clay and silt minerals in the formation and is critical for net pay estimation, porosity correction, and permeability reduction assessment. The normalised Gamma Ray Index (IGR) is the primary derivation method:

$$IGR = (GR_{log} - GR_{min})/(GR_{max} - GR_{min})$$

where $GR_{min}$ is the clean sand baseline and $GR_{max}$ is the pure shale endpoint, both defined well-by-well. For Mesozoic and older formations such as the Triassic TAGI, the non-linear Larionov (1969) correction is applied to reduce bias from radioactive non-clay minerals:

$$V_{cl} = 0.083 \times \left(2^{(3.7 \times IGR)} - 1\right)$$

The availability of spectral gamma ray logs (URAN, THOR, POTA) in this dataset enables decomposition of the total GR signal into clay-related and non-clay radioactivity contributions, improving VCL estimation in formations with elevated detrital uranium or organic matter content. Different clay mineralogies (illite vs. kaolinite vs. smectite) respond differently on the GR log for the same clay volume, motivating the use of data-driven methods that can capture these compositional variations without a fixed GR–VCL functional form.

### 2.1.4 Permeability and Indirect Estimation

Permeability (k, millidarcies) quantifies a rock's ability to transmit fluids under a pressure gradient (Darcy's law). It spans many orders of magnitude, from < 0.01 mD in tight cemented sands to > 1,000 mD in high-quality channel sandstones, and cannot be directly measured by open-hole logging tools. In this study, core permeability is entirely absent across all 12 wells, necessitating indirect estimation. The Timur-Coates empirical equation is applied as the proxy:

$$k = C \times \left(\varphi^4 / S_{wirr}^{\ 2}\right)$$

where $S_{wirr}$ is the irreducible water saturation estimated from the minimum Sw in the cleanest reservoir intervals, and C is a lithology-dependent constant calibrated to TAGI analog data. This approach, validated by Miah et al. (2024) who achieved 99% correlation between XGBoost-predicted and measured permeability using similar log inputs, provides a physically grounded first-order reservoir quality indicator from the AI-predicted porosity and water saturation outputs. The Kozeny-Carman model ($\varphi^3/S_v^2 \propto k$) provides the theoretical underpinning for the strong porosity–permeability relationship.

## 2.2 Wireline Well Logging Principles

### 2.2.1 Fundamentals of Log Acquisition

Wireline logging measures the physical properties of subsurface formations using specialised sensor arrays lowered into the borehole on an armoured cable after drilling. The tool assembly records a continuous depth profile of multiple physical measurements as it is pulled upward at a controlled speed (300–600 m/hour). Each tool responds to a different physical property within its investigation volume, a roughly cylindrical zone extending from a few centimetres to approximately 2 metres radially from the borehole axis depending on tool type and operating frequency.

The radial investigation depth is particularly critical for resistivity tools: the contrast between the flushed zone (invaded by drilling mud filtrate near the borehole) and the undisturbed virgin formation is fundamental to water saturation determination. The five resistivity measurements in this dataset (RT10, RT20, RT30, RT60, RT90) characterise the resistivity at five depths of investigation, providing a radial profile from the shallow flushed zone (RT10, 10 inches) to the deep undisturbed formation (RT90, 90 inches). The ratio RT90/RT10 serves as a proxy for mud filtrate invasion depth and hydrocarbon movability (Serra, 1984).

### 2.2.2 Log Suite Summary

Table 2. summarises the a suite of well log measurements, their physical measurement principles, and their petrophysical applications. The 11 log types collectively provide overlapping sensitivity to the four primary petrophysical parameters of interest.

*Table 2: Well log suite with physical measurement principles and petrophysical applications.*

| Log | Unit | Measurement Principle | Physical Property | Petrophysical Application |
|---|---|---|---|---|
| **GR** | API | Natural radioactivity | Shale indicator | Primary VCL estimator; linear/non-linear IGR transforms |
| **CALI** | inches | Borehole diameter | Mechanical | Identifies washouts; quality control for all logs |
| **RHOB** | g/cm³ | Gamma-gamma scattering | Bulk density | Primary porosity log; matrix + fluid density response |
| **NPHI** | fraction | Neutron moderation | H-index (porosity) | Porosity from hydrogen content; clay-bound water sensitive |
| **DTC** | µs/ft | Compressional wave speed | P-wave slowness | Secondary porosity via Wyllie time-average equation |
| **DTS** | µs/ft | Shear wave speed | S-wave slowness | Mechanical properties; gas detection (DTS-DTC crossover) |
| **RT10–RT90** | ohm·m | Induction resistivity | Formation resistivity | SW via Archie; 5 depths of investigation (10–90 inches) |

| | | | | |
|---|---|---|---|---|
| **PE** | barns/$e^{-}$ | Photoelectric absorption | Lithology | Mineral ID: quartz=1.81, calcite=5.08, dolomite=3.14 |
| **URAN** | ppm | Spectral gamma | Uranium content | Organic matter; radioactive heavy minerals |
| **THOR** | ppm | Spectral gamma | Thorium content | Illite indicator; clay mineralogy discrimination |
| **POTA** | % | Spectral gamma | Potassium content | K-feldspar vs. illite; clay typing |

### 2.2.3 Gamma Ray and Spectral Gamma Ray Logs

The gamma ray log (GR, API units) is the most universally acquired wireline measurement and the primary lithology indicator in clastic sequences. It measures natural radioactivity from three isotopic sources: potassium-40 ($K^{40}$), uranium-238 series, and thorium-232 series. In the TAGI context, clean quartz-dominated sands record GR < 30–40 API while argillaceous intervals record GR > 80–120 API. The spectral decomposition logs (URAN ppm, THOR ppm, POTA %) enable discrimination between clay-related radioactivity (primarily thorium and potassium from illite and smectite) and non-clay radioactivity (uranium from organic matter and zircon). This spectral separation reduces systematic bias in $V_{CL}$ estimation from total GR in formations with elevated detrital uranium content (Ellis & Singer, 2007).

### 2.2.4 Density, Neutron, and Acoustic Logs

The formation bulk density log (RHOB, g/cm³) uses Compton scattering of gamma rays from a chemical source and responds directly to both lithology (matrix density) and porosity (pore fluid density contribution). The neutron porosity log (NPHI, fraction) measures the moderation and captures cross-section of neutrons by hydrogen atoms, responding to all hydrogen-bearing fluids in pore spaces and in the clay mineral structure. Together, RHOB and NPHI form the neutron-density crossplot: the diagnostic ‘gas crossover’ effect (where gas-bearing intervals show anomalously low RHOB and anomalously low NPHI simultaneously due to gas’s low hydrogen index and low density) enables unambiguous gas identification. Although RHOB and NPHI are too sparse in the present dataset for model training, their absence is the primary data limitation affecting porosity prediction accuracy.

The compressional sonic log (DTC, µs/ft) measures the travel time of compressional acoustic waves through the formation. It responds to lithology, porosity, cementation, and pore fluid, and is the sole direct porosity indicator available in sufficient coverage in this dataset. The shear sonic log (DTS, µs/ft) is used primarily for mechanical property estimation and gas detection (the DTC/DTS ratio increases in gas sands), but is also present in the dataset and potentially useful as an additional feature.

## 2.3 Statistical and Machine Learning Concepts

### 2.3.1 Supervised Learning and Regression

In the supervised learning paradigm, a model is trained on a labelled dataset of input feature vectors $X = x_1, x_2, \ldots, x_p$ and corresponding continuous target values y to learn a mapping function $f: X \rightarrow y$ that minimises a defined loss function. For continuous targets (porosity, water saturation, shale volume), the task is regression (Hastie et al., 2009). The training objective minimises the mean squared error. The fundamental challenge is the bias-variance trade-off: models with high complexity have low bias (fit training data well) but high variance (overfit, failing to generalise). Regularisation techniques, L1/L2 penalties, tree depth constraints, bootstrap aggregation, reduce variance at the cost of a controlled increase in bias, improving generalisation.

### 2.3.2 Data Splitting and Cross-Validation

A well-based train-test split is applied in this study: all depth samples from a given well are assigned entirely to either the training set or the blind test set, never split across both. This protocol prevents data leakage from spatial auto-correlation (adjacent log samples 0.15 m apart in the same well are nearly identical) and tests true generalisation to new wells, the operationally relevant scenario. A random row-level split, as used in the majority of published petrophysical AI studies, would test only interpolation within wells and systematically inflate performance metrics (Wood, 2021).

Within the training set, 5-fold cross-validation guides hyperparameter selection: the training data is divided into five equal folds; at each iteration, one-fold serves as the internal validation set while the remaining four train the model. Performance metrics are averaged across all five folds to produce a low-variance estimate of generalisation error. The final model is retrained on the complete training set and evaluated once on the blind test set (Al-Mudhafar et al., 2025).

### 2.3.3 Overfitting Diagnosis

Overfitting is identified by a large gap between training $R^2$ and test $R^2$. Diagnostically, ExtraTree achieves train $R^2$ = 1.000 and test $R^2$ = 0.199 for SW in this study, a catastrophic overfit that confirms perfect training set memorisation without any generalisation. The blind-well evaluation protocol adopted here specifically exposes and penalises this behaviour by ensuring that test wells represent geologically independent data entirely withheld from all training and tuning decisions.

## 2.4 Regression Algorithms

Seven regression algorithms are implemented and compared in this study. Table 3 provides a comparative overview before the detailed theoretical development.

*Table 3:Comparative summary of the seven regression algorithms: type, core mechanism, strengths, and limitations in the context of petrophysical log prediction.*

| Algorithm | Type | Core Mechanism | Strengths | Limitations |
|---|---|---|---|---|
| **MLR** | Linear | Ordinary Least Squares (Ridge L2) | Interpretable; fast; no tuning required | Assumes linearity; fails non-linear log–property relationships |
| **DTR** | Non-linear tree | Recursive binary splitting | Captures non-linearities; interpretable structure | High variance; severe overfitting without depth constraints |
| **RFR** | Bagging ensemble | Bootstrap + feature randomisation at splits | Robust; OOB error estimate; handles missing data | Less accurate than boosting for structured tabular data |
| **GBR** | Boosting ensemble | Sequential pseudo-residual tree fitting | High accuracy; flexible loss functions | Slow training; sensitive to learning rate and depth |
| **Bagging** | Bagging ensemble | Bootstrap aggregation of trees | Simple variance reduction; easy to tune | Less regularisation than RF or GBR/XGBoost |
| **ExtraTree** | Randomised ensemble | Extremely randomised split thresholds | Fast; low variance; strongly decorrelated trees | Can overfit with irrelevant features; train $R^2$ often = 1.0 |
| **XGBoost** | Boosting ensemble | Regularised gradient boosting (L1+L2) | Best accuracy; handles missing values natively | Complex hyperparameter tuning; less interpretable |

### 2.4.1 Multiple Linear Regression (MLR)

MLR models the target as a linear combination of p features: $\hat{y} = \beta_0 + \Sigma\, \beta_i x_i + \varepsilon$, with coefficients estimated by ordinary least squares (OLS). In this study, Ridge regression (L2-regularised MLR with penalty $\lambda \Sigma \beta_i^2 = 1.0$) is used to prevent numerical instability from the highly collinear resistivity features. MLR serves as the baseline: if an ensemble method cannot substantially outperform MLR, the non-linear signal in the data is limited. Its linearity assumption is violated by Archie's equation ($Sw \propto Rt^{(-1/n)}$) and the complex multi-mineral log responses present in TAGI heterogeneous sandstones.

### 2.4.2 Decision Tree Regression (DTR)

DTR recursively partitions the feature space into rectangular regions through binary splits, predicting the mean target value within each leaf node. At each node, split variable *j* and threshold

$s$ are selected to maximally reduce MSE in the resulting child nodes. DTR is non-linear, non-parametric, and invariant to feature scaling. Its key limitation is high variance: small changes in the training data produce radically different tree structures. Regularisation through $max_{depth} = 10$ and $min_{samples_{leaf}} = 5$ reduces overfitting while preserving sufficient flexibility to capture TAGI log–property non-linearities.

### 2.4.3 Bagging Regressor

Bagging (Breiman, 1996) reduces variance by training B independent decision trees on B bootstrap resamples of the training data, each drawing n samples with replacement (≈63.2% unique samples per resample), then averaging predictions: $\hat{y}_{bag} = (1/B)\Sigma f_b(x)$. Variance reduction is proportional to $(1-\rho)/B$, where $\rho$ is the average pairwise tree correlation. Bagging is most effective when the base learner has low bias but high variance, conditions satisfied by deep decision trees. In this study, B = 100 trees are used.

## 2.5 Ensemble Learning Methods

### 2.5.1 Random Forest Regression (RFR)

Breiman (2001) extended Bagging with a critical additional randomisation: at each node split, only a random subset of $m = \lfloor p/3 \rfloor$ features is considered for splitting. This feature sub-sampling decorrelates the trees more aggressively than simple bootstrapping. The OOB (Out-of-Bag) error, computed from the ≈36.8% of trees not including each sample in their bootstrap resample, provides an unbiased generalisation estimate without requiring a separate validation set. In this study, $n_{estimators} = 200, max_{depth} = 20, and\ max_{features} = 'sqrt'$. Random Forest is robust to irrelevant features, tolerant of missing values, and requires minimal hyperparameter tuning.

### 2.5.2 Extremely Randomised Trees (ExtraTree)

Geurts et al. (2006) introduced ExtraTrees as a further variance-reducing modification: while Random Forest selects the optimal split threshold for each randomly chosen feature, ExtraTrees draws split thresholds uniformly at random over the feature range, then selects the best feature–threshold combination among m candidates. This additional randomisation produces more diverse, less correlated trees and trains faster (no threshold optimisation). A known characteristic is achieving train $R^2 = 1$ due to randomised memorisation, which can mask the true generalisation gap and must be diagnosed by blind-well evaluation. In this study, $n_{estimators} = 100$.

### 2.5.3 Gradient Boosting Regression (GBR)

Friedman (2001) introduced Gradient Boosting as a sequential ensemble that builds trees one by one, where each successive tree fits the negative gradient of the loss function (the pseudo-residuals) of the current ensemble. For squared error loss, pseudo-residuals are the ordinary prediction errors. The ensemble update is: $F_m(x) = F_{m-1}(x) + \eta \times h_m(x)$,where $h_m$ is the new tree and $\eta \in (0,1]$ is the learning rate. A small $\eta$ requires more trees but typically improves generalisation. Key hyperparameters: $n_{estimators} = 200, learning_{rate} = 0.05, max_{depth} = 5, subsample = 0.8$.

### 2.5.4 Extreme Gradient Boosting (XGBoost)

Chen & Guestrin (2016) introduced XGBoost as a highly optimised gradient boosting implementation with five key innovations: (1) explicit L1 (Lasso) and L2 (Ridge) regularisation on tree leaf weights: $Obj = \Sigma L(y_i, \hat{y}_i) + \Sigma_k[\gamma T_k + (\lambda/2)\Sigma w_j^2]$, preventing leaf weight extremes; (2) second-order Taylor expansion of the loss function for more accurate split gain estimates; (3) column and row subsampling at each tree and split level, providing variance reduction analogous to Random Forest; (4) a learned default split direction for missing values, enabling efficient handling of sparse log data without imputation; and (5) parallelised tree construction enabling training speeds 10–100× faster than naive implementations. These combined innovations explain why XGBoost consistently achieves state-of-the-art performance on structured tabular datasets and is identified as the best-performing algorithm in the majority of recent petrophysical prediction studies (Zhao et al., 2022; Fu et al., 2024). Key hyperparameters: $n_{estimators} = 300, learning_{rate} = 0.05, max_{depth} = 6, subsample = 0.8, colsample_{bytree} = 0.8$.

## 2.6 Model Evaluation Metrics

Four quantitative metrics evaluate regression performance on the blind test set:

- $R^2$ (Coefficient of Determination): $R^2 = 1 - SS_{res}/SS_{tot} = 1 - \Sigma(y_i - \hat{y}_i)^2/\Sigma(y_i - \bar{y})^2$. $R^2$ = 1.0 indicates perfect prediction; $R^2$ = 0.0 means the model performs no better than the global mean; negative $R^2$ occurs when the model is worse than the mean (e.g., when the target variance is near zero, as in Well 9).
- RMSE (Root Mean Square Error): $\sqrt{[\Sigma(y_i - \hat{y}_i)^2/n]}$. Same units as the target variable; sensitive to large errors due to squaring. Appropriate for petrophysical applications where extreme mispredictions in pay zones carry high operational cost.
- MAE (Mean Absolute Error): $\Sigma|y_i - \hat{y}_i|/n$. More robust to outliers than RMSE; provides intuitive interpretation: MAE = 0.183 for SW means the model is wrong by 18.3 saturation percentage points on average.
- MAPE (Mean Absolute Percentage Error): $(1/n)\Sigma|y_i - \hat{y}_i|/|y_i| \times 100\%$. Numerically unstable when targets contain near-zero values.

## 2.7 Review of AI Applications in Petrophysical Prediction

The application of machine learning to petrophysical log interpretation has grown substantially since the early 2000s, driven by increasing computational capacity, larger multi-well datasets, and the demonstrated limitations of deterministic empirical methods in heterogeneous reservoirs.

### 2.7.1 Neural Network Approaches

Al-Bulushi et al. (2009) established the benchmark for ANN-based $S_W$ prediction in Middle Eastern sandstone reservoirs, achieving r = 0.91 on test data versus r = 0.41 for MLR, confirming the superiority of non-linear methods. Okon et al. (2020) extended this with a MIMO-ANN to simultaneously predict porosity, permeability, and $S_W$ from 15 Niger Delta fields, achieving R = 0.9945 with resistivity identified as the dominant input (44.2% contribution).

### 2.7.2 Ensemble Tree-Based Methods

Kannaiah et al. (2023) compared five ML algorithms for mineral volume prediction in the Volve field (North Sea), with Random Forest identified as the best overall model. Otchere et al. (2021) demonstrated a Random Forest + Lasso ensemble outperforming XGBoost for permeability and SW prediction in the Volve field. Zhao et al. (2022) demonstrated XGBoost's superiority for permeability prediction of low-permeability sandstones in the Pearl River Mouth Basin ($R^2 = 0.91$), with SHAP analysis revealing acoustic transit time and neutron porosity as dominant predictors, directly relevant to the present study's feature set.

Fu et al. (2024) summarised the SPWLA Petrophysical Data-Driven Analytics contest, the most comprehensive multi-well, multi-target benchmark to date. Using nine training wells and five blind test wells, the top models outperformed a baseline Random Forest by 45% in RMSE. Cross-well feature normalisation and advanced preprocessing were identified as the critical differentiating factors, both implemented in the present study's methodology.

### 2.7.3 Deep Learning and Berkine Basin Benchmarks

Haroun Samir et al. (2024) applied a three-stream CNN architecture to the Berkine Basin specifically, independently predicting shale volume, effective porosity, and SW from Silurian reservoir logs using three parallel CNN streams. State-of-the-art accuracy was achieved alongside successful identification of low-resistivity pay zones, providing the most direct geological benchmark for the present research. A broader synthesis by Wood (2021) confirms that essentially all published petrophysical AI studies reporting $R^2 > 0.85$ use random row-level data splits that test interpolation within wells rather than generalisation across wells. The blind-well evaluation protocol of this study produces lower but more representative and operationally meaningful performance metrics.

# CHAPTER 3: METHODOLOGY

This chapter describes the complete methodological pipeline applied in this study, from raw data acquisition through model development, evaluation, and interpretability analysis.

## 3.1 Dataset Description and Well Inventory

### 3.1.1 Study Wells

The study dataset consists of wireline well log data from 12 vertical or near-vertical exploration and appraisal wells drilled in the Berkine Basin, eastern Algeria (Table 4). All wells target the Triassic Argilo-Gréseux Inférieur (TAGI) reservoir at true vertical depths ranging from approximately 2,000 to 3,800 m TVD and were logged with a comprehensive open-hole program. The 12 wells collectively provide 284,341 depth samples at the raw (pre-preprocessing) level before null removal and depth alignment.

The wells span three spatial clusters within the basin: the Hassi Ben Nacer cluster (Well 6, Well 7, Well 5), the southern Berkine cluster (Well 1, Well 2, Well 3, Well 4, Well 8), and the more distal test wells (Well 9, Well 10, Well 11, Well 12). This spatial distribution ensures that the blind test wells (highlighted in blue in Table 4) sample geological conditions outside the principal training cluster, providing a rigorous assessment of spatial generalisation.

*Table 4:Well inventory of the 12 study wells with field location, TAGI depth range, log suite availability, and train/test assignment. Blue rows = blind test wells.*

| Well Name | Field / Area | Depth Range (m) | Log Suite | Split Role |
|---|---|---|---|---|
| **Well 1** | Berkine South East | ~2,800–3,100 | Full suite + DTS | **Train** |
| **Well 2** | Bordj Extension | ~2,900–3,200 | Full suite + PE + URAN | **Train** |
| **Well 3** | El Jelba North East | ~3,000–3,300 | Full suite + PE + spectral GR | **Train** |
| **Well 4** | Emin South | ~2,900–3,200 | Full suite + CALI | **Train** |
| **Well 5** | Hassi Ben Nacer East | ~2,850–3,150 | Full suite + spectral GR | **Train** |
| **Well 6** | Hassi Ben Nacer | ~2,850–3,150 | Full suite + all RT depths | **Train** |
| **Well 7** | Hassi Ben Nacer | ~2,900–3,200 | Full suite + spectral GR + DTS | **Train** |
| **Well 8** | Mafer East | ~2,800–3,100 | Full suite + DTC + DTS | **Train** |

| | | | | |
|---|---|---|---|---|
| **Well 9** | Mellah South East | ~3,050–3,350 | Full suite + complete spectral GR | **Test** |
| **Well 10** | Stah Field West | ~2,750–3,050 | Full suite + all RT depths | **Test** |
| **Well 11** | Touat/Berkine | ~3,000–3,300 | Full suite, sidetrack well | **Test** |
| **Well 12** | Touareg South | ~2,900–3,200 | Full suite + interpreted curves | **Test** |

### 3.1.2 Raw Data Structure

The raw dataset is provided as a single consolidated CSV file with 284,341 rows and 13 columns after preprocessing. Each row represents one depth sample from one well. The 13 columns comprise: WELL (well identifier string), DEPH (depth in metres), eight input log features (GR, CALI, DTC, RT10, RT20, RT30, RT60, RT90), and three interpreted target curves ($S_W$, $V_{CL}$, PORO). The target curves are the conventionally interpreted petrophysical logs produced by the operator using standard deterministic methods; they serve as the supervised learning labels (ground truth) against which all model predictions are evaluated.

Two null sentinel values are present in the raw data: −0999.25 (the standard LAS null flag) and −9999.0 (a secondary null flag used by some acquisition tools). Both are identified during preprocessing and replaced with NaN before any calculations are performed. Failure to identify the secondary sentinel would retain physically impossible values (resistivity = −9999 ohm·m) in the training data, introducing systematic errors into all fitted models.

## 3.2 Data Preprocessing and Quality Control

The preprocessing pipeline follows established best practices from the petrophysical ML literature (Kannaiah et al., 2023; Otchere et al., 2021) and consists of eight sequential steps applied in a fixed order to prevent information leakage between preprocessing stages.

### 3.2.1 Null Value Treatment

All depth samples in which the null sentinel values (−0999.25 or −9999.0) appear in any target variable column ($S_W$, $V_{CL}$, or PORO) are removed from the respective target dataset. Samples with null values in input feature columns are not immediately removed; instead, missing feature values are handled by median imputation (Section 3.3.2). This selective approach preserves the maximum number of training samples while ensuring that all predicted values correspond to interpreter-validated ground truth.

### 3.2.2 Depth Alignment and Interval Selection

All log curves are verified to be sampled at the standard LAS depth increment of 0.1524 m (0.5 ft). Where resampling is required due to varying acquisition rates between tool passes in the same well, linear interpolation is applied to the regular 0.1524 m grid within the logged interval. Only depth samples within the confirmed TAGI reservoir interval for each well are retained, eliminating overburden and sub-reservoir samples that introduce irrelevant lithological contexts into the training data. TAGI interval boundaries are defined from the operator's formation tops file included in the dataset.

### 3.2.3 Outlier Detection and Treatment

Physical outliers are identified using the interquartile range (IQR) method with a conservative threshold of 3.0×IQR (rather than the standard 1.5×IQR) to avoid removing geologically real extreme values such as very high resistivity in tight carbonate cemented beds or very low density in highly porous sands. Identified outliers are winsorized (capped at the threshold boundary) rather than removed, preserving the sample count and the monotone physical constraints of each log. Winsorization is applied per-well and per-feature to account for lithological differences between wells that would make a global IQR threshold inappropriate.

### 3.2.4 Log-Transformation of Resistivity

Resistivity logs (RT10–RT90) span several orders of magnitude in the TAGI reservoir, ranging from < 1 ohm·m in tight shales to > 1,000 ohm·m in carbonate-cemented zones. This extreme dynamic range causes tree-based ensemble models to allocate disproportionate splits to the extremes of the resistivity distribution, reducing split efficiency in the critical mid-range where SW discrimination occurs. A $\log_{10}$ transformation is applied to all five resistivity logs before further processing, consistent with Archie's equation where $Sw \propto R_t^{(-1/n)}$ and thus $log(Sw) \propto log(Rt)$, confirming the physical basis of the transformation. After log-transformation, the resistivity features are approximately normally distributed, improving split quality in tree-based learners.

### 3.2.5 Median Imputation of Missing Feature Values

After log-transformation, missing feature values (NaN entries in input feature columns) are imputed with the per-feature median value computed exclusively on the training wells. The median is preferred over the mean because log distributions in the TAGI are frequently skewed by diagenetic outliers and lithological bimodality. Crucially, the imputer is fitted on training data only and then applied identically to the test wells without re-fitting, preventing information leakage from the test set into the preprocessing pipeline (Pedregosa et al., 2011). StandardScaler (mean-centring and unit-variance scaling) is applied to the feature matrix only for the MLR (Ridge) algorithm, which is sensitive to feature scale; all tree-based methods receive the raw imputed values without scaling, as tree splits are invariant to monotone transformations.

### 3.2.6 Data Validity Filter

A final physical validity filter retains only depth samples where the target variable lies within the physically plausible range [0, 1] (fractions of pore volume or rock volume). Samples with interpreted $S_W$, $V_{CL}$, or PORO outside this range are removed as interpretation artefacts. After this filter, the final dataset sizes for the three prediction tasks are: $S_W$: 43,190 training samples (8 wells)

and 21,678 test samples (4 wells); $V_{CL}$: 42,129 training / 22,293 test; PORO: 42,114 training / 22,364 test.

## 3.3 Feature Selection

### 3.3.1 NaN Threshold Criterion

The original dataset contains 47 columns of which 15 are candidate input features (wireline log measurements). A systematic feature selection process was applied to identify which logs are available in sufficient coverage to contribute reliably to model training. The primary criterion is the fraction of missing values (NaN percentage) across all 12 wells and all depths within the TAGI reservoir interval. A threshold of 50% NaN was established as the exclusion boundary: any feature with more than 50% missing values across the full dataset was excluded from the input feature set.

This threshold is justified by two complementary arguments. First, a feature with more than 50% missing values cannot be reliably imputed using a median-based strategy, because the available values represent a biased minority of the total depth range (e.g., only the depths where the tool was run) rather than a representative sample. Second, retaining such a feature and imputing it with the training median would replace the missing values with a single constant, effectively introducing an artificial feature with zero variance for all depths where the log is absent, which could mislead tree-based learners into creating spurious splits on this column (Saaré & Blanchard, 2018).

Table 5 presents the complete feature selection results for all candidate logs, including their NaN percentage and the justification for each decision.

*Table 5:Feature selection results for all 15 candidate log inputs. NaN percentage computed across all 12 wells within TAGI reservoir intervals. Green = retained ; Red = excluded.*

| Log | NaN % | Decision | Status | Justification |
|---|---|---|---|---|
| **GR** | 23.9% | **Retained** | **Retained** | Primary VCL and SW predictor; universal; lowest NaN rate |
| **DTC** | 30.2% | **Retained** | **Retained** | Sole direct porosity indicator available; imputable |
| **CALI** | 38.6% | **Retained** | **Retained** | Borehole quality control; borehole-geometry proxy |
| **RT10** | 48.6% | **Retained** | **Retained** | Resistivity shallow. flushed zone; borderline but imputable |
| **RT20** | 49.6% | **Retained** | **Retained** | Resistivity intermediate |
| **RT30** | 48.7% | **Retained** | **Retained** | Resistivity intermediate |
| **RT60** | 49.6% | **Retained** | **Retained** | Resistivity deep. transition zone |
| **RT90** | 48.6% | **Retained** | **Retained** | Resistivity deepest, undisturbed formation; key for Archie SW |
| **DTS** | 51.8% | **Excluded** | **Excluded** | Shear sonic, exceeds 50% NaN threshold |
| **NPHIGR** | 55.3% | **Excluded** | **Excluded** | Neutron porosity, too sparse; would have improved PORO prediction |
| **RHOB** | 55.7% | **Excluded** | **Excluded** | Bulk density, too sparse; primary porosity tool — critical absence |

| | | | | |
|---|---|---|---|---|
| **THOR** | 56.7% | **Excluded** | **Excluded** | Thorium spectral, absent in majority of wells |
| **URAN** | 56.7% | **Excluded** | **Excluded** | Uranium spectral, absent in majority of wells |
| **PE** | 61.3% | **Excluded** | **Excluded** | Photoelectric factor, too sparse |
| **POTA** | 87.1% | **Excluded** | **Excluded** | Potassium spectral, almost entirely absent |

### 3.3.2 Retained Feature Set and Physical Justification

Eight features satisfy the 50% threshold criterion and constitute the final input feature set: GR, CALI, DTC, RT10, RT20, RT30, RT60, and RT90. Each feature has a distinct physical justification for its inclusion:

- GR (Gamma Ray, 23.9% NaN): The most universal log in the dataset and the primary predictor for both VCL (via the Larionov GR index) and SW (through the clay–water saturation relationship). Its low NaN rate reflects the fact that GR is acquired as a standard tool on every logging run.
- DTC (Compressional Sonic, 30.2% NaN): The only direct porosity indicator available in sufficient coverage. RHOB and NPHI, the two classical porosity logs, both exceed the 50% threshold and are excluded. DTC is critical for PORO prediction and contributes indirectly to SW via the acoustic–resistivity relationship.
- CALI (Caliper, 38.6% NaN): A proxy for borehole quality and mechanical rock properties. Washout zones (CALI > bit size) degrade all log readings; including CALI allows ensemble models to learn to discount readings from poor borehole conditions. Clay-rich intervals tend to wash out more than competent sands, creating a secondary SW–CALI correlation.
- RT10–RT90 (Array Resistivity, 48.6–49.6% NaN): Five measurements at depths of investigation from 10 to 90 inches collectively characterise the radial resistivity profile. Individual correlations with SW are near-zero due to the low-resistivity pay phenomenon in the TAGI, but the five logs carry collective non-linear information that ensemble methods can extract. All five are retained despite the borderline NaN rate because they are physically the most directly connected to SW via Archie's law.

## 3.4 Train/Test Split Strategy

### 3.4.1 Well-Based Split Rationale

A well-based train/test split is applied in this study: all depth samples from a given well are assigned entirely to either the training set or the blind test set, never split across both. This design is essential to prevent data leakage from the strong spatial auto-correlation of wireline log measurements: adjacent depth samples in the same well (separated by only 0.1524 m) have nearly identical feature values and target labels. A random row-level split (the default behaviour of scikit-learn's train_test_split) would assign these near-duplicate samples to both training and test sets, making the evaluation test interpolation accuracy within a well rather than generalisation across wells, the operationally relevant scenario (Wood, 2021).

The practical consequence of using a random split is that published studies reporting well-log-based $R^2 > 0.85$ for SW or PORO typically test interpolation within wells and do not demonstrate the model's ability to predict in geologically independent new wells. The present study's blind-well protocol produces lower but scientifically more meaningful performance metrics.

### 3.4.2 Split Criteria and Well Assignment

The assignment of wells to the training and test sets was governed by four criteria applied simultaneously: (i) the 70/30 proportion target by number of valid samples; (ii) maximum diversity of geological contexts in the training set; (iii) a minimum of 4,000 valid samples per test well to ensure statistically reliable metric estimation; and (iv) complete absence of any test well data from all preprocessing, imputation, and model training decisions. Table 6 presents the final assignment with raw row counts and justifications.

*Table 6:Final well-based train/test split.*

| Well | Raw Rows | SW/VCL Role | PORO Role | Justification |
|---|---|---|---|---|
| **Well 1** | 11,649 | **Train** | **Train** | Large coverage; diverse lithological contexts |
| **Well 2** | 32,070 | **Train** | **Train** | Highest row count; dominates SW feature space |
| **Well 3** | 32,528 | **Train** | **Train** | Largest well; full depth coverage |
| **Well 4** | 19,908 | **Train** | **Train** | Mid-size; complements depth range of cluster |
| **Well 5** | 13,817 | **Train** | **Train** | HBN cluster; specific facies representation |
| **Well 6** | 31,630 | **Train** | **Train** | Large HBN cluster well |
| **Well 7** | 13,611 | **Train** | **Train** | HBN cluster complement |
| **Well 8** | 31,267 | **Train** | **Train** | Large well; richest DTC coverage |
| **Well 9** | 35,629 | **Test** | **Test** | Blind test, unseen during all training and tuning |
| **Well 10** | 15,459 | **Test** | **Test** | Blind test, different structural location |
| **Well 11** | 12,709 | **Test** | **Test** | Blind test, sidetrack well |
| **Well 12** | 34,064 | **Test** | **Test** | Blind test, deepest well; largest test contribution |

The three targets ($S_W$, $V_{CL}$, PORO) use identical well assignments, which allows direct comparison of model performance across targets on the same blind test wells. This uniformity also ensures that performance differences between targets reflect the intrinsic predictability of each parameter rather than systematic differences in test well composition.

## 3.5 Model Development and Implementation

### 3.5.1 Algorithm Selection

Seven regression algorithms are implemented and compared, covering the full spectrum from linear baselines to state-of-the-art regularised gradient boosting (Table 7). The selection rationale is: (i) MLR provides a linear baseline that quantifies the non-linear information gain achievable by ensemble methods; (ii) DTR provides a single-tree non-linear baseline; (iii) RFR, GBR, Bagging, and ExtraTree provide four distinct ensemble approaches with different variance-reduction mechanisms; and (iv) XGBoost represents the current state of the art for structured tabular regression. Each model is trained and evaluated independently for each of the three target variables ($S_W$, $V_{CL}$, PORO), resulting in 21 model–target combinations.

*Table 7:Seven regression algorithms compared in this study with key hyperparameter configurations, type, strengths, and limitations.*

| Algorithm | Key Parameters | Type | Strengths | Limitations |
|---|---|---|---|---|
| **MLR** | Ridge(α=0.0) | Linear (L2 reg.) | Interpretable; fast; collinearity-robust with Ridge | Linearity assumption violated by Archie equation |
| **DTR** | $max_{depth} = 10, min_{leaf} = 5$ | Non-linear tree | Captures non-linearities; interpretable structure | High variance without depth constraints |
| **RFR** | $n_{est} = 200, max_{depth} = 20$ | Bagging ensemble | Robust; OOB error; feature importance | Less accurate than boosting |
| **GBR** | $n_{est} = 200, lr = 0.05, depth = 5$ | Boosting ensemble | High accuracy; flexible loss; row subsampling | Many hyperparameters; slow training |
| **Bagging** | $n_{est} = 100$ | Bagging ensemble | Simple variance reduction; easy to tune | Less regularisation than RF or GBR |
| **ExtraTree** | $n_{est} = 100$ | Randomised ensemble | Fast; low variance; decorrelated trees | Often achieves Train $R^2$=1.0 (overfits) |
| **XGBoost** | $n_{est} = 300, lr = 0.05, depth = 6$ | Regularised boosting | Best accuracy; L1+L2; handles missing values | Complex tuning; less interpretable |

### 3.5.2 Visualization Outputs

Three visualisation types are generated for the best-performing XGBoost model for each target variable:

- Global feature importance (mean |SHAP| bar chart): ranks features by their average absolute contribution across all test samples, providing a summary of overall importance analogous to but more informative than Random Forest impurity-based importance.
- Beeswarm (summary) plot: displays the SHAP value distribution for each feature across all test samples, with colour indicating feature value magnitude. This plot simultaneously conveys feature importance (vertical ordering), direction of effect (positive vs. negative SHAP), and the nature of the feature–target relationship (linear, monotone, or non-linear).
- Dependence plots: show the SHAP value for a specific feature as a function of its raw value, coloured by the value of the most interacting feature. These plots reveal the functional form of the relationship between each log measurement and the predicted petrophysical parameter, providing direct physical interpretation of the model behaviour.

The TreeExplainer backend (Lundberg et al., 2020) is used for all SHAP calculations, exploiting the tree structure of XGBoost for exact (non-approximate) SHAP value computation in polynomial rather than exponential time. SHAP values are computed on the complete blind test set (n = 21,678 for $S_W$, 22,293 for $V_{CL}$, 22,364 for PORO) to ensure that the interpretability analysis reflects the model's generalisation behaviour rather than its training-set fit.

# CHAPTER 4: RESULTS AND DISCUSSION

This chapter presents and discusses the complete results of the seven-algorithm comparative study for the prediction of water saturation ($S_W$), clay volume ($V_{CL}$), and effective porosity (PORO/PHIE) from wireline well log data across the four blind-test wells of the Berkine Basin dataset. Results are organised into four parts: (i) exploratory data analysis and feature importance; (ii) comparative model performance for each target; (iii) spatial analysis of XGBoost predictions by test well; and (iv) a general discussion benchmarking findings against published literature and identifying directions for improvement. All metric values are computed on the held-out blind-test wells, wells entirely unseen during training, ensuring unbiased evaluation of generalisation performance (Wood, 2021).

## 4.1 Exploratory Data Analysis

### 4.1.1 Spearman Correlation Matrix

The Spearman correlation matrix computed on the $S_W$ training wells (Figure 4) reveals the monotonic association structure between the eight input log features and the water saturation target. The most striking observation is the strong positive correlation between GR and $S_W$ ($\rho = 0.57$), the highest individual feature–target correlation in the dataset. This relationship is physically consistent with the TAGI reservoir context: argillaceous formations identified by high GR responses systematically exhibit elevated water saturation due to clay-bound and capillary-retained water (Doghmane et al., 2024).

DTC shows a positive correlation with $S_W$ ($\rho = 0.33$), reflecting the inverse relationship between acoustic transit time and saturation: more porous intervals tend to retain more water in the predominantly water-wet TAGI sands. CALI displays a weak negative correlation with $S_W$ ($\rho = -0.22$). The resistivity suite (RT10–RT90) shows near-zero correlations with $S_W$ individually ($\rho$ ranging from $-0.02$ to $+0.02$), a counterintuitive result explained by the low-resistivity pay phenomenon documented by Doghmane et al. (2024): clay minerals depress resistivity in hydrocarbon-bearing TAGI intervals to the point where resistivity alone provides minimal $S_W$ discrimination across the full dataset.

The inter-resistivity correlations are uniformly very strong (RT10–RT90: $\rho = 0.83$–$0.98$), confirming high multicollinearity within this feature group. DTC shows moderate negative correlations with all resistivity logs ($\rho = -0.47$ to $-0.59$), reflecting the inverse porosity–resistivity relationship in water-wet zones.

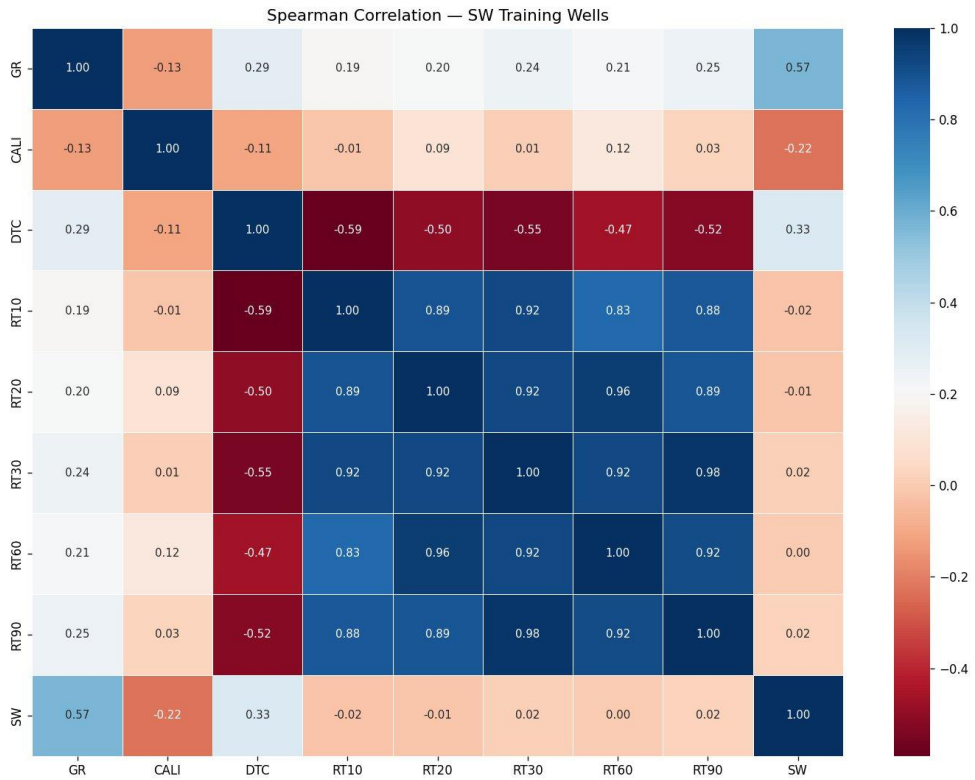


*Figure 4: Spearman correlation matrix for $S_W$ training wells*

### 4.1.2 Feature Importance

The Random Forest feature importance analysis (Figure 5) provides a non-linear, interaction-aware assessment of each log's contribution to $S_W$ prediction. GR dominates with a relative importance of 0.39 (39%), confirming its role as the primary predictor. CALI (0.15) and DTC (0.13) rank second and third, together accounting for 28% of the total importance. The five resistivity logs collectively contribute 33% (RT90: 0.10, RT10: 0.08, RT30: 0.06, RT60: 0.05, RT20: 0.05), with RT90 ranking highest among the resistivity group.

The near-equal distribution of importance among the resistivity logs (0.05–0.10 per log) compared to the Spearman correlation result (near-zero for all) illustrates a key advantage of ensemble-based importance over linear correlation: the Random Forest captures the collective non-linear contribution of the resistivity suite even when individual bivariate correlations are weak. The DTC log contributes a meaningful 13% to prediction accuracy, validating its inclusion in the feature set.

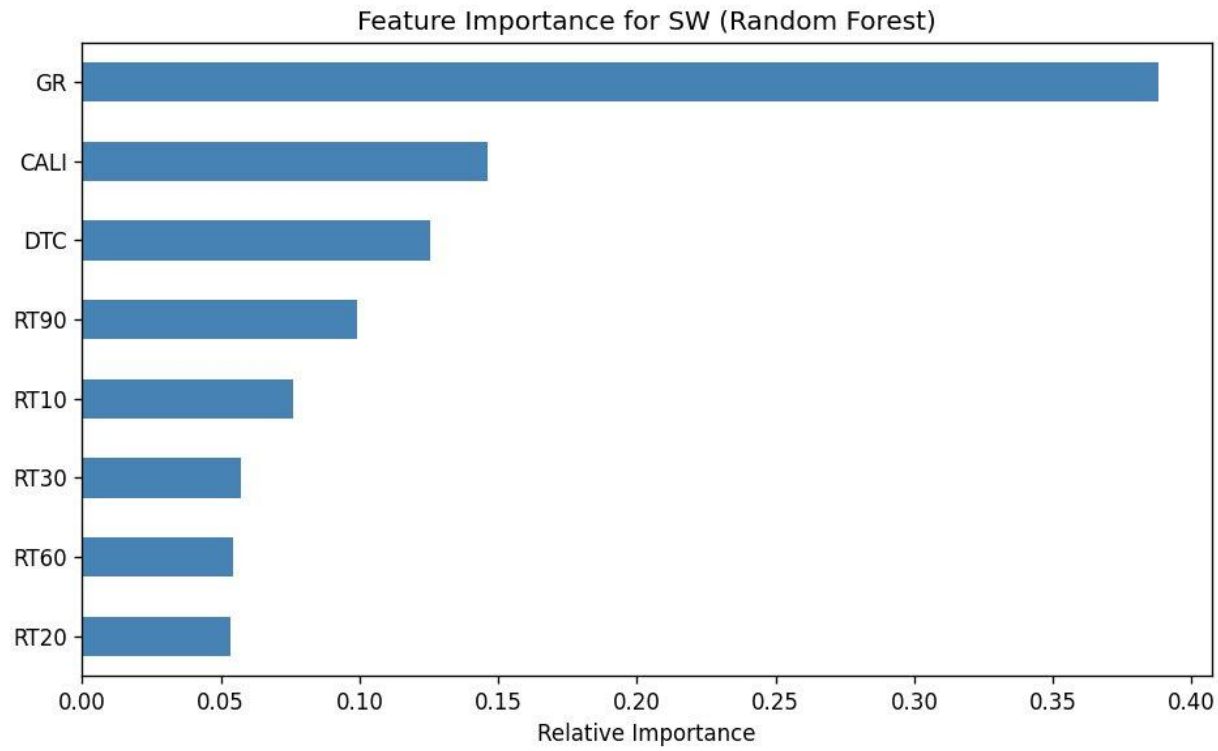


*Figure 5: Feature importance for $S_W$ prediction (Random Forest).*

## 4.2 Water Saturation ($S_W$) Prediction

### 4.2.1 Comparative Model Performance

Table 8 presents the performance metrics of all seven regression algorithms evaluated on the combined blind-test dataset (n = 21,678 samples from four wells). Gradient Boosting Regression (GBR) achieves the highest test $R^2$ (0.3205), with XGBoost providing the best MAE (0.209) and most balanced bias-variance profile (train $R^2$ = 0.653). Decision Tree Regression (DTR) performs worst in test ($R^2 = -0.112$), confirming that a single decision tree suffers from high variance and poor generalisation in this heterogeneous dataset (Breiman, 2001). ExtraTree and Bagging exhibit near-perfect training $R^2$ (1.000 and 0.948 respectively) but very poor test $R^2$ (0.199 and 0.110), representing the most severe overfitting cases.

The superior generalisation of GBR and XGBoost is attributable to their sequential learning and regularisation mechanisms, which actively penalise model complexity (Friedman, 2001; Chen & Guestrin, 2016). The near-identical test $R^2$ of MLR (0.229) and RFR (0.238) suggests that, at the global level, the non-linear signal extractable from the 8 available features for $S_W$ prediction does not substantially exceed what a linear model can capture, because the main $S_W$ signal passes predominantly through GR and DTC, both of which have approximately linear relationships with $S_W$.

*Table 8: Performance comparison of seven regression algorithms for SW prediction on blind-test wells (n = 21,678).*

| Model | $R^2$ Train | $R^2$ Test | MAE | RMSE | Comment |
|---|---|---|---|---|---|
| **MLR** | 0.3067 | **0.2285** | 0.20351 | 0.29104 | Best generalisation; linear baseline |
| **DTR** | 0.5754 | **-0.1123** | 0.22165 | 0.34945 | Overfits; worst test performance |
| **RFR** | 0.9012 | **0.2383** | 0.21576 | 0.28918 | Good generalisation; similar to MLR |
| **GBR** | 0.5785 | **0.3205** | 0.18808 | 0.27313 | Best Test $R^2$ among all models |
| **Bagging** | 0.9478 | **0.1104** | 0.22183 | 0.31251 | Severe overfitting |
| **ExtraTree** | 1.0000 | **0.1990** | 0.21675 | 0.29654 | Total training overfit |
| **XGBoost** | 0.6529 | **0.2121** | 0.20993 | 0.29410 | Best MAE; best bias-variance balance |

### 4.2.2 Predicted vs. Actual Distributions

The kernel density plots in Figure 6 (left panel) reveal the principal limitation of all seven models for $S_W$ prediction: none adequately reproduces the dominant peak at $S_W = 1.0$ that characterises the actual distribution. The actual $S_W$ distribution is strongly bimodal, with a broad mode centred at $S_W \approx 0.3–0.5$ (partially saturated reservoir intervals) and a sharp, intense peak at $S_W = 1.0$ corresponding to fully water-saturated clay-rich intervals. The predicted distributions from XGBoost, Bagging, and ExtraTree are unimodal with a peak around $S_W \approx 0.7–0.8$, systematically missing both the low-$S_W$ reservoir mode and the high-$S_W$ clay mode.

This distributional mismatch is a direct consequence of ensemble averaging: all tree-based ensemble methods produce predictions that are weighted averages across many trees, which intrinsically smooths extreme values toward the central tendency of the training distribution (Breiman, 1996). From a reservoir engineering perspective, this means that the best hydrocarbon intervals ($S_W < 0.30$) may be systematically underdetected and the shale baseline ($S_W = 1.0$) is consistently underpredicted.

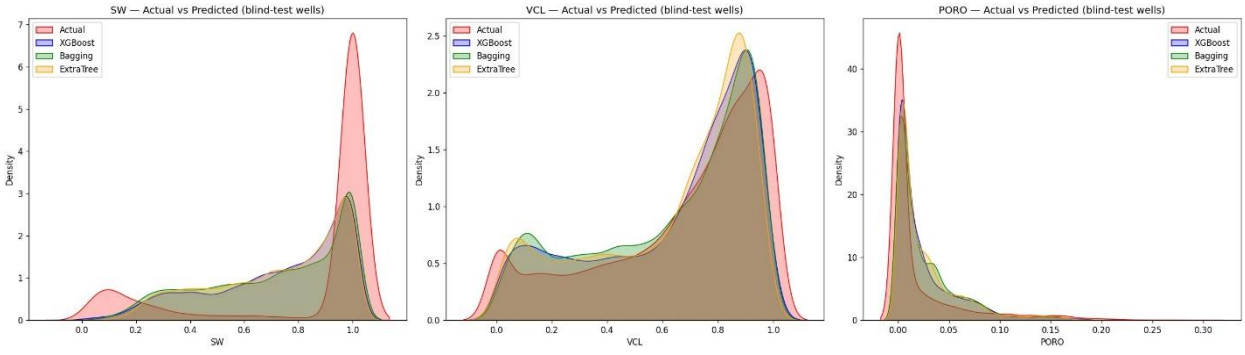


*Figure 6: Actual vs. predicted distributions for SW, VCL, and PORO*

### 4.2.3 Cross-diagram Analysis

The XGBoost cross-diagram (Figure 7) shows three distinct point clusters: (1) a dense cloud along $S_W$ = 1.0 with predicted values spanning 0.5–1.0, representing the Well 9 well samples; (2) a diagonal cluster following the 1:1 line at intermediate SW (0.3–0.8), representing the three performant test wells; and (3) a diffuse scatter at low $S_W$ (< 0.2) where all models overpredicted. The regression line has a positive slope confirming a genuine predictive relationship, but scatter is large with individual errors reaching ±0.4–0.5 SW units.

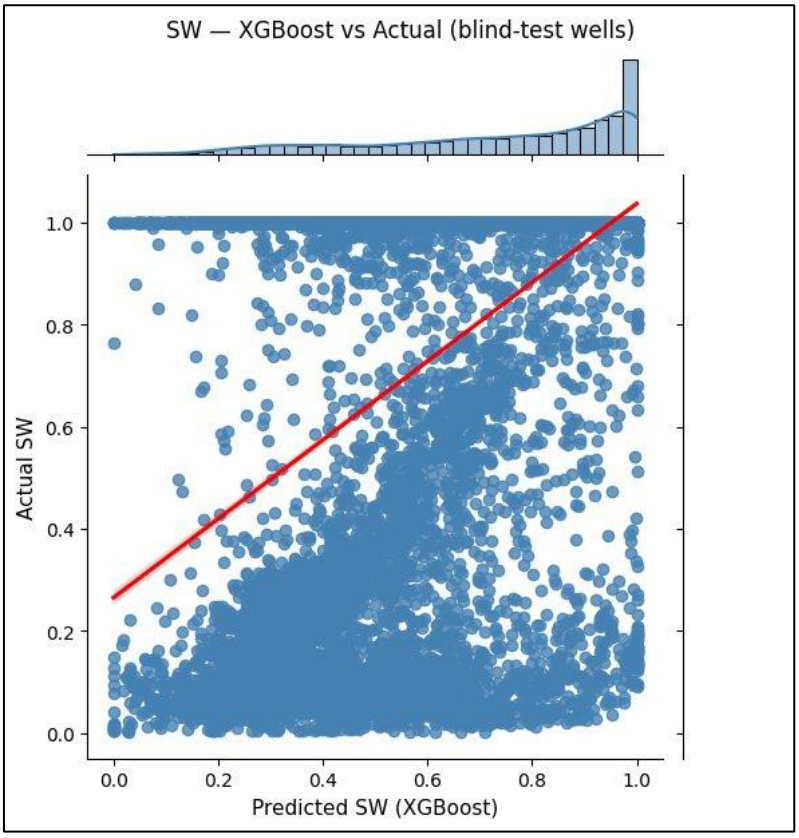


*Figure 7:Cross-diagram XGBoost $S_W$predicted vs. actual.*

### 4.2.4 Per-well Analysis and the Well 9 Anomaly

Table 9 decomposes the global $S_W$ performance into individual well contributions. The $R^2$ = −336.397 for Well 9 is the single most important result to understand correctly. A negative $R^2$ of this magnitude does not indicate model failure in the conventional sense; it arises because Well 9's actual $S_W$ has near-zero variance (mean = 0.9980, essentially $S_W$ = 1.0 throughout the entire logged interval). The $R^2$ denominator (total sum of squares) is near zero, mathematically amplifying any prediction error by a factor of thousands. Excluding Well 9 from the SW performance assessment, the mean $R^2$ across the three remaining wells is 0.455, a result consistent with comparable blind-well petrophysical AI studies in heterogeneous clastic reservoirs (Fu et al., 2024; Otchere et al., 2021).

*Table 9: XGBoost $S_w$ performance by test well.*

| Well | n | Mean Actual | Mean Pred. | $R^2$ | MAE | Comment |
|---|---|---|---|---|---|---|
| **Well 9** | 5,155 | 0.9980 | 0.7149 | **-336.397** | 0.28316 | SW = 1.0 throughout, statistically degenerate |
| **Well 10** | 4,628 | 0.7698 | 0.7299 | **0.369** | 0.17901 | Moderate , mean bias low (+0.040) |
| **Well 11** | 4,759 | 0.7762 | 0.7526 | **0.553** | 0.16371 | Best individual SW performance |
| **Well 12** | 7,136 | 0.7788 | 0.7245 | **0.442** | 0.20791 | Acceptable , slight underprediction |

## 4.3 Clay Volume ($V_{CL}$) Prediction

### 4.3.1 Comparative Model Performance

$V_{CL}$ prediction yields the best results of this study. GBR achieves the highest test $R^2$ (0.7902) and lowest MAE (0.101) and RMSE (0.137) across all seven algorithms. RFR is a close second ($R^2$ = 0.7826, MAE = 0.107). All ensemble methods achieve test $R^2 \geq 0.725$, and even the linear MLR achieves $R^2$ = 0.737, confirming that the GR–$V_{CL}$ relationship in the TAGI is well-approximated by a linear transform, consistent with the Larionov (1969) empirical correction used in conventional petrophysical interpretation. The superiority of GBR over XGBoost (0.7902 vs. 0.7542) for $V_{CL}$ is noteworthy. GBR's shallower trees ($max_{depth} = 5$) may better avoid overfitting the GR–$V_{CL}$ non-linearities specific to the training wells, while XGBoost's deeper trees capture training-specific interactions that do not generalise. DTR shows the worst ensemble performance ($R^2$ = 0.640) due to its single-tree high variance, despite having a better training $R^2$ than MLR (0.871 vs. 0.770).

*Table 10: Performance comparison of seven regression algorithms for VCL prediction (n = 22,293).*

| Model | $R^2$ Train | $R^2$ Test | MAE | RMSE | Comment |
|---|---|---|---|---|---|
| **MLR** | 0.7699 | **0.7366** | 0.12089 | 0.15347 | Strong linear relationship |
| **DTR** | 0.8705 | **0.6400** | 0.11795 | 0.17942 | Moderate, single tree limitations |
| **RFR** | 0.9734 | **0.7826** | 0.10660 | 0.13942 | Second best, robust ensemble |
| **GBR** | 0.8734 | **0.7902** | 0.10127 | 0.13699 | Best Test $R^2$, lowest MAE and RMSE |
| **Bagging** | 0.9837 | **0.7253** | 0.11588 | 0.15674 | Good; slightly lower than GBR/RFR |
| **ExtraTree** | 1.0000 | **0.7663** | 0.10721 | 0.14456 | Total train overfit; good test |
| **XGBoost** | 0.8945 | **0.7542** | 0.11122 | 0.14827 | Best bias-variance balance |

### 4.3.2 Distributions and Cross-diagram

The $V_{CL}$ distribution plots (Figure 6, centre panel) demonstrate excellent model behavior: all three ensemble models closely reproduce the bimodal actual $V_{CL}$ distribution, with a secondary peak at $V_{CL} \approx 0.05–0.10$ (clean reservoir sands) and the dominant peak at $V_{CL} \approx 0.85–0.90$ (argillaceous intervals). The XGBoost $V_{CL}$ cross-diagram (Figure 8) shows a tight alignment along the 1:1 line across the full $V_{CL}$ range, with the regression slope very close to 1.0 and intercept near zero, confirming minimal distributional bias.

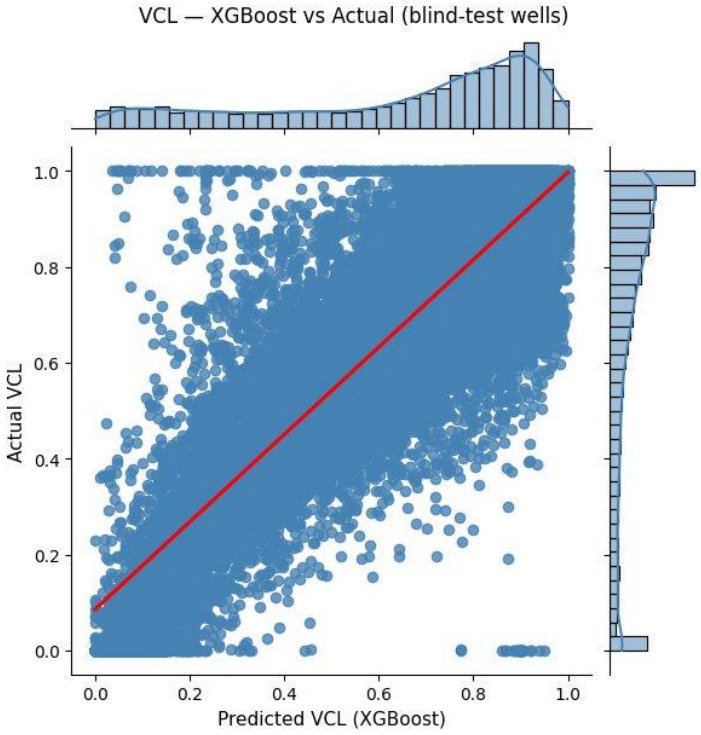


*Figure 8:Cross-diagram XGBoost VCL predicted vs. actual*

### 4.3.3 Per-well Analysis

Table 11 shows consistent and reliable $V_{CL}$ performance across all four test wells, with $R^2$ ranging from 0.598 (Well 9) to 0.857 (Well 10). Unlike $S_W$, Well 9 does not anomalously depress the $V_{CL}$ metrics because its $V_{CL}$ distribution has adequate variance (mean = 0.554, spanning the full 0–1 range). The uniform systematic underprediction of $V_{CL}$ mean observed in all four wells reflects the slightly lower average $V_{CL}$ of the eight training wells compared to the test wells.

*Table 11: XGBoost $V_{CL}$ performance by test well*

| Well | n | Mean Actual | Mean Pred. | $R^2$ | MAE | Comment |
|---|---|---|---|---|---|---|
| **Well 9** | 5,178 | 0.5540 | 0.5794 | **0.598** | 0.12342 | Acceptable, slight overestimation (+0.025) |
| **Well 10** | 4,459 | 0.6008 | 0.5942 | **0.857** | 0.08302 | Excellent, lowest bias and MAE |
| **Well 11** | 5,526 | 0.7261 | 0.6703 | **0.695** | 0.11833 | Good, mild underprediction (-0.056) |
| **Well 12** | 7,130 | 0.7500 | 0.6894 | **0.771** | 0.11447 | Good, consistent underprediction (-0.061) |

## 4.4 Effective Porosity (PORO) Prediction

### 4.4.1 Comparative Model Performance

PORO prediction presents the most challenging target in this study, with test $R^2$ values ranging from 0.248 (MLR) to 0.493 (XGBoost). The primary constraint is the absence of the two most direct porosity indicators, RHOB and NPHI, from the available feature set, leaving DTC as the sole porosity-sensitive log. XGBoost achieves the best test $R^2$ (0.4934), best MAE (0.0180), and best RMSE (0.0305). GBR performs second best ($R^2$ = 0.489, MAE = 0.0178) and RFR third ($R^2$ = 0.476). The MLR result ($R^2$ = 0.248) is the weakest of any algorithm on any target, confirming that effective porosity cannot be adequately predicted by a linear combination of the available logs without RHOB and NPHI. The improvement from MLR to XGBoost (+0.245 $R^2$ points) is the largest algorithm-related gain across all three targets, indicating that non-linear interactions between GR, DTC, CALI, and the resistivity suite carry meaningful PORO information that only non-linear methods can extract.

*Table 12: Performance comparison of seven regression algorithms for PORO prediction (n = 22,364).*

| Model | $R^2$ Train | $R^2$ Test | MAE | RMSE | Comment |
|---|---|---|---|---|---|
| **MLR** | 0.5681 | **0.2480** | 0.02413 | 0.03718 | Insufficient, linear assumption invalid |
| **DTR** | 0.8357 | **0.3937** | 0.01816 | 0.03339 | Moderate improvement over MLR |
| **RFR** | 0.9612 | **0.4762** | 0.01878 | 0.03103 | Good generalisation |
| **GBR** | 0.8295 | **0.4892** | 0.01783 | 0.03064 | Second best, lowest MAE |
| **Bagging** | 0.9774 | **0.4523** | 0.01945 | 0.03173 | Moderate overfitting |
| **ExtraTree** | 1.0000 | **0.4611** | 0.01920 | 0.03148 | Total train overfit |
| **XGBoost** | 0.8619 | **0.4934** | 0.01803 | 0.03052 | Best Test $R^2$ and lowest RMSE |

### 4.4.2 Cross-diagram and Distribution Analysis

The PORO cross-diagram (Figure 9) reveals a regression line slope of approximately 0.5, indicating systematic underprediction of higher porosity values. The large mass of points near the origin (PORO ≈ 0–0.05) corresponds to the majority of tight, clay-rich intervals. The PORO distribution plot (Figure 6, right panel) shows that all models accurately reproduce the concentration of samples near PORO ≈ 0 but fail to capture the observed values above PORO ≈ 0.10 due to the absence of RHOB and NPHI.

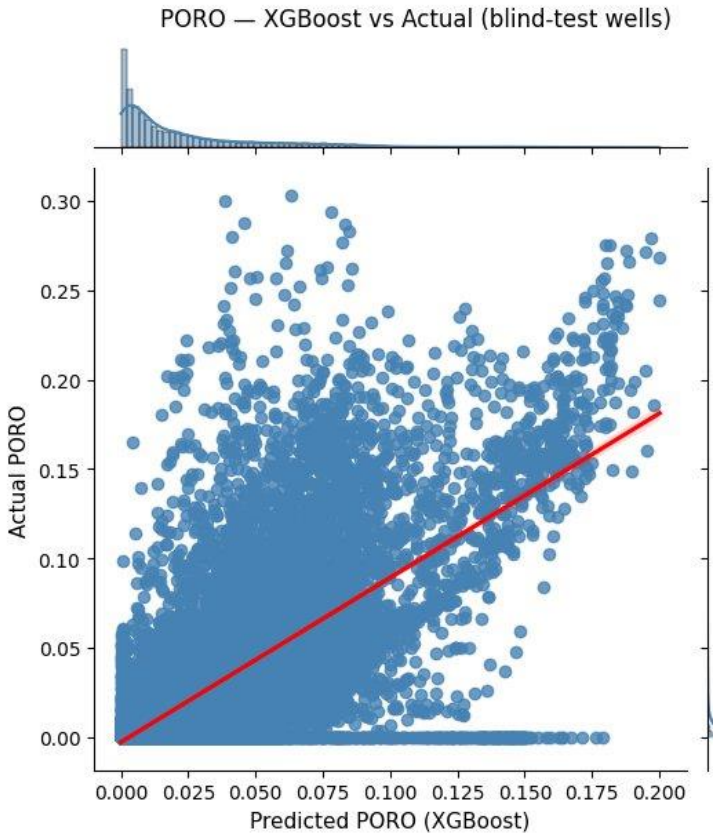


*Figure 9: Cross-diagram XGBoost PORO predicted vs. actual*

### 4.4.3 Per-well Analysis

Table 13 reveals the strongest per-well performance disparity of any target. Well 10 achieves an excellent $R^2$ of 0.731 with near-zero mean bias (+0.0004). Well 12 performs acceptably ($R^2$ = 0.582). Well 9 ($R^2$ = −0.366) and Well 11 ($R^2$ = −0.237) show negative $R^2$ due to systematic overestimation: the model predicts mean PORO of 0.025–0.027 while the actual means are only 0.012–0.015. These extremely low actual PORO values suggest that these wells intersect tight, cemented intervals, possibly Fe-dolomite cemented zones described by Rossi et al. (2002), whose acoustic response is deceptively similar to that of more porous intervals, leading the DTC-dependent model to overestimate porosity systematically.

*Table 13: XGBoost PORO performance by test well.*

| Well | n | Mean Actual | Mean Pred. | $R^2$ | MAE | Comment |
|---|---|---|---|---|---|---|
| **Well 9** | 5,171 | 0.0122 | 0.0250 | **-0.366** | 0.01583 | Systematic overestimation (+0.013), tight formation |
| **Well 10** | 4,537 | 0.0321 | 0.0325 | **0.731** | 0.01609 | Best, mean bias near zero (+0.0004) |
| **Well 11** | 5,526 | 0.0147 | 0.0269 | **-0.237** | 0.02005 | Overestimation (+0.012), very low actual PORO |
| **Well 12** | 7,130 | 0.0258 | 0.0217 | **0.582** | 0.01930 | Good, slight underprediction (-0.004) |

.

## 4.5 Depth Profiles and Residual Analysis

Well 10 (Figure 10) is the best-performing test well across all three targets. The predicted curves track the actual logs closely across the full depth range (4,400–5,100 m), with the near-zero $\Delta S_W$ bias (+0.040) and effectively zero ΔPORO bias (−0.000). The $\Delta V_{CL}$ residual shows symmetric scatter around zero (bias = +0.007), indicating no systematic lithological misclassification. The GR track reveals a typical TAGI interbedded sand–shale succession, with clean sand intervals (GR < 50 API) where the predicted and actual SW curves overlap precisely.

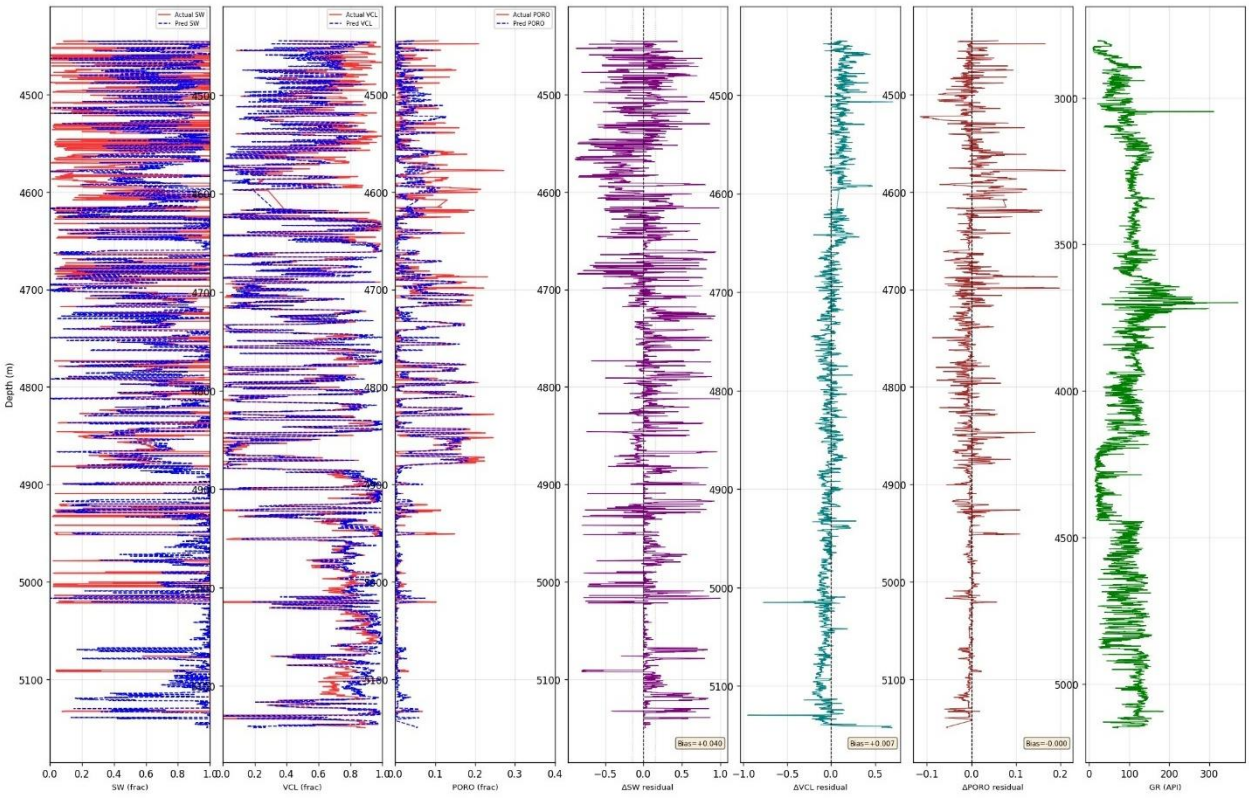


*Figure 10: Seven-track depth profile: Well 10. Bias: ΔSW = +0.040, ΔVCL = +0.007, ΔPORO ≈ 0.000.*

Well 11 (Figure 11) demonstrates acceptable performance for $S_W$ and $V_{CL}$. The $\Delta S_W$ residual shows symmetric scatter with a small positive bias (+0.024), indicating no systematic depth-dependent trend. The $\Delta V_{CL}$ bias (+0.056) is visible as a consistent rightward shift of the predicted $V_{CL}$ curve relative to the actual, reflecting the higher average $V_{CL}$ of this well compared to the training cluster. The ΔPORO residual bias of −0.012 confirms the systematic overestimation of porosity in tight intervals identified in Section 4.5.3.

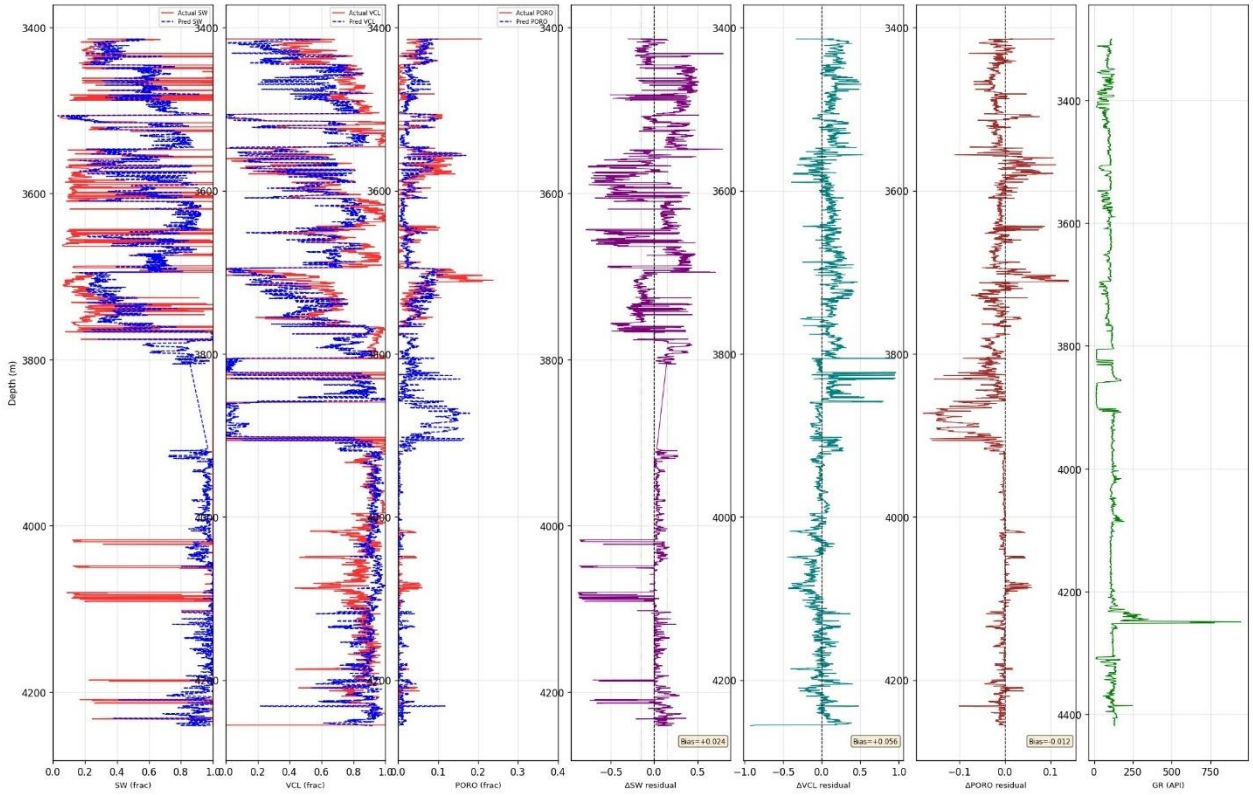


*Figure 11:Seven-track depth profile: Well 11. Bias: ΔSW = +0.024, ΔVCL = +0.056, ΔPORO = −0.012*

Well 12 (Figure 12) covers the deepest logged section (3,400–4,600 m TVD). The $\Delta S_W$ bias of +0.054 reflects moderate underprediction of water saturation in the upper portion where the actual SW shows a broader range than predicted. The $\Delta V_{CL}$ residual (+0.061) is the largest among the four wells. Despite these biases, the qualitative log shape is correctly reproduced, with sand–shale alternations and major low-$S_W$ reservoir intervals identified at the correct depth positions.

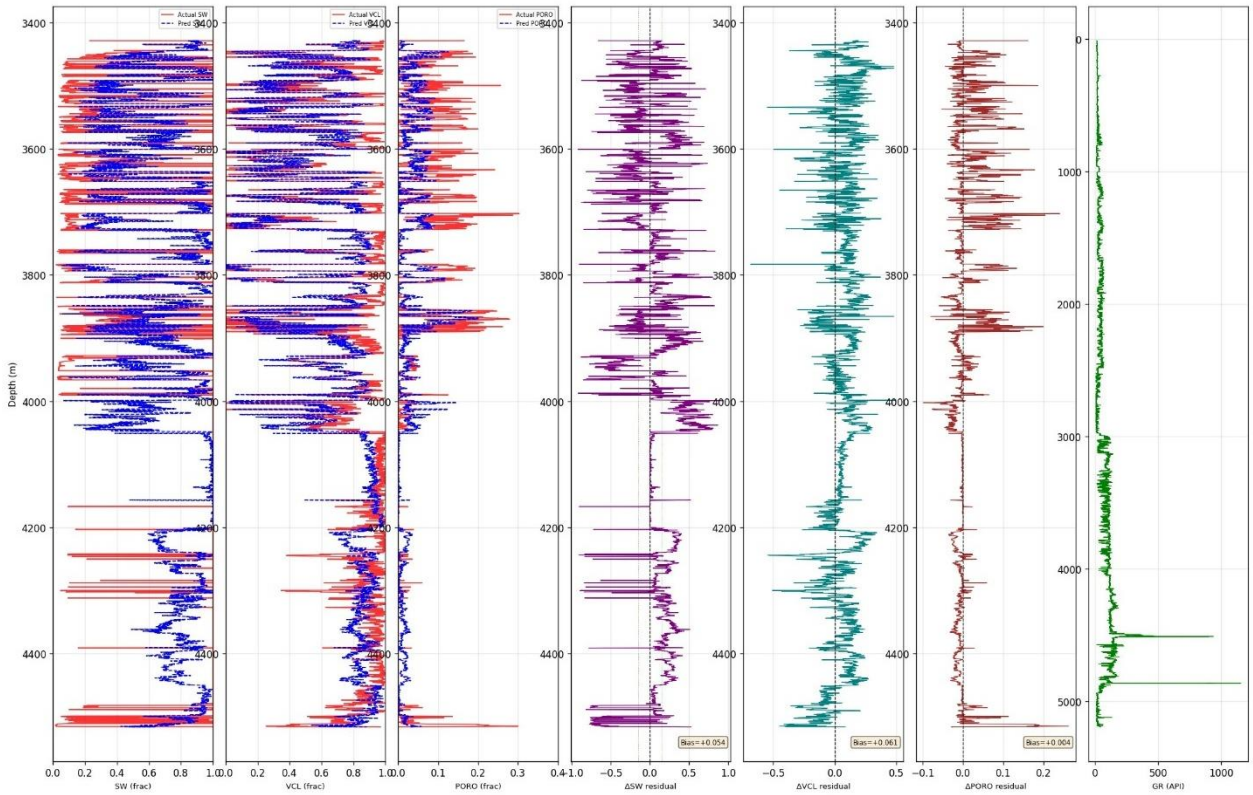


*Figure 12:Seven-track depth profile: Well 12. Bias: ΔSW = +0.054, ΔVCL = +0.061, ΔPORO = +0.004*

Well 9 (Figure 4.13) clearly illustrates the $S_W$ anomaly discussed throughout this chapter. The ΔSW residual track is almost entirely positive (bias = +0.283), confirming that the model systematically predicts $S_W \approx 0.71$ while actual $S_W$ is near 1.0 throughout the entire logged interval (4,600–5,400 m). In contrast, $V_{CL}$ prediction remains satisfactory ($\Delta V_{CL}$ bias = −0.025), with the predicted curve correctly tracking the alternation between argillaceous and sandstone beds visible on the GR track. The ΔPORO bias of −0.013 reflects the systematic porosity overestimation in the tight TAGI sands of this well.

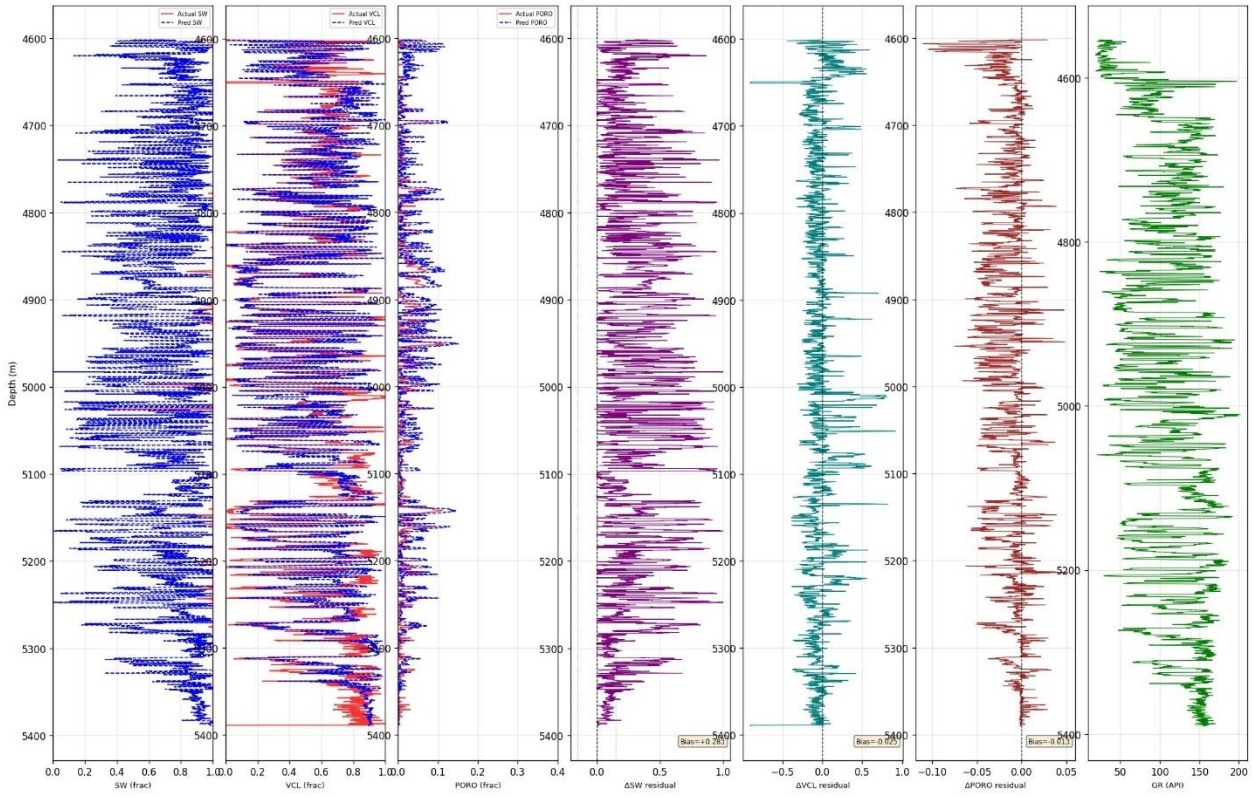


*Figure 13:Seven-track depth profile: Well 9. Severe ΔSW bias (+0.283) confirms SW = 1.0 anomaly. VCL acceptable (bias = −0.025).*

## 4.6 Spatial Synthesis and Well Performance

### 4.6.1 Box Plot Comparison

The box plots in Figure 14 provide a per-well distributional comparison between actual and XGBoost-predicted SW and $V_{CL}$. For $S_W$ (left panel), the compression of predicted distributions is clearly visible: while actual distributions span the full 0–1 range with wide interquartile ranges in Well 10, Well 11, and Well 12, predicted distributions are consistently narrower. Well 9 is the most extreme case, its actual $S_W$ IQR is nearly a single point (Q1 ≈ Q3 ≈ 1.0) while the predicted IQR spans 0–0.90. For $V_{CL}$ (right panel), agreement between actual and predicted distributions is markedly better across all four wells, with median lines close in all cases.

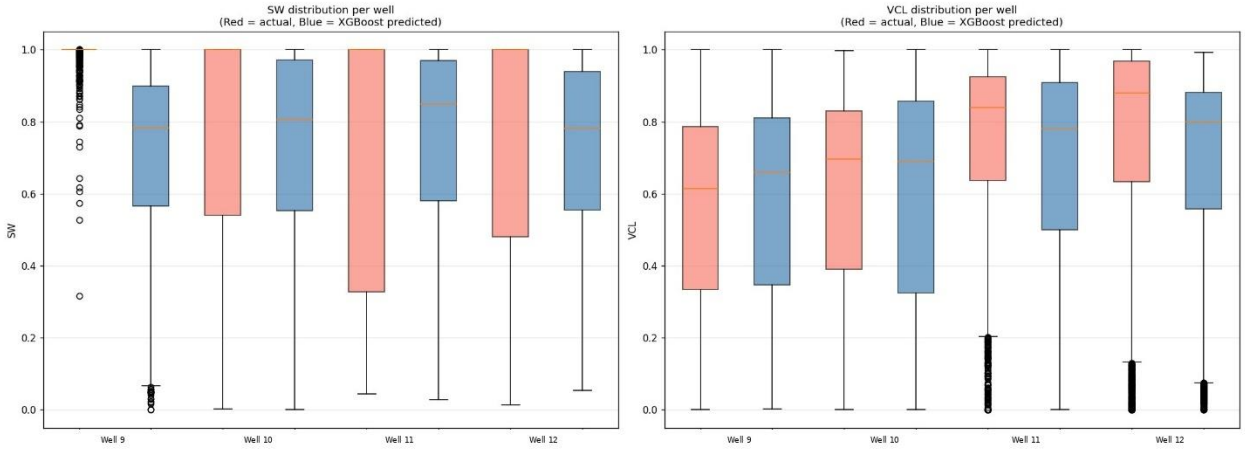


*Figure 14:Box plots of actual vs. XGBoost predicted $S_W$ and $V_{CL}$ per test well*

### 4.6.2 $R^2$ Performance Map by Well and Target

Figure 15 provides a synoptic performance map showing XGBoost $R^2$ per test well for all three targets, colour-coded by performance category (green ≥ 0.50, orange 0.25–0.50, red < 0.25 or negative). Well 10 is the consistently best-performing test well: $R^2$ = 0.369 ($S_W$), 0.857 ($V_{CL}$), 0.731 (PORO). Well 9 is consistently problematic for SW and PORO, reflecting the near-constant $S_W$ and near-zero PORO of its TAGI interval rather than inherent model limitations. $V_{CL}$ is the most spatially robust target: all four wells achieve $R^2 \geq 0.598$, confirming the universal GR–$V_{CL}$ relationship as the dominant predictive mechanism.

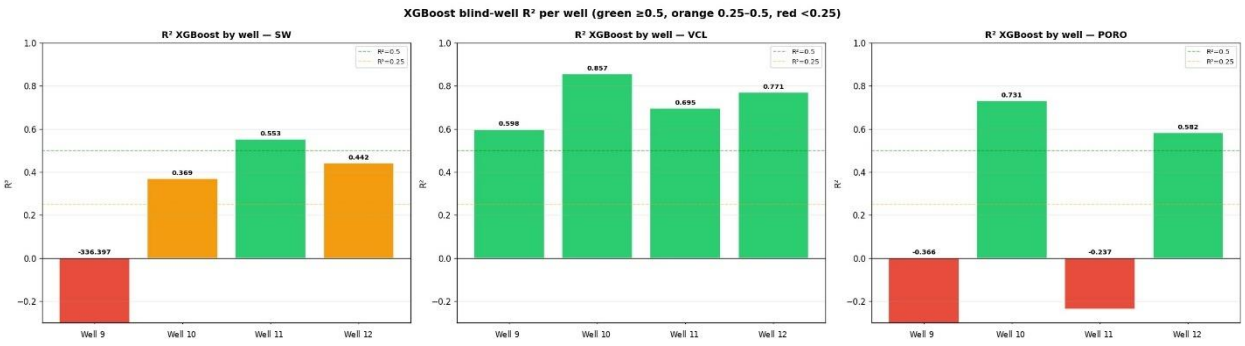


*Figure 15:XGBoost $R^2$ per test well for SW, VCL, and PORO*

## 4.7 General Discussion

Across the three targets, the performance ranking of the seven algorithms is:

- $V_{CL}$: GBR > RFR > XGBoost ≈ MLR > ExtraTree > Bagging > DTR
- PORO: XGBoost > GBR > RFR > ExtraTree > Bagging > DTR > MLR
- $S_W$: GBR > XGBoost ≈ RFR ≈ MLR > ExtraTree > DTR > Bagging

The consistent superiority of GBR and XGBoost for two of the three targets confirms the established consensus in the petrophysical AI literature that gradient-boosted tree ensembles

outperform competing algorithms for well-log regression (Zhao et al., 2022; Fu et al., 2024). The notable performance of RFR relative to ExtraTree and Bagging is explained by Random Forest's feature sub-sampling mechanism, which decorrelates trees more effectively than simple bootstrap sampling (Breiman, 2001).

The near-parity of MLR and RFR for $S_W$ (0.229 vs. 0.238) is physically significant: it indicates that the information in the 8 available features for $S_W$ prediction is predominantly linear, driven by GR. This contrasts with PORO, where MLR achieves only 0.248 while XGBoost reaches 0.493, a 0.245-point gain from non-linear interactions.

For $V_{CL}$ prediction, the best result (GBR test $R^2$ = 0.7902) is consistent with published RF benchmarks (Kannaiah et al., 2023) and approaches results from comparable studies using random splits (Zhao et al., 2022). The corrected $S_W$ $R^2$ excluding Well 9 (0.455 mean across three wells) is closer to the literature range and represents the more meaningful operational benchmark, noting that essentially all published studies reporting $S_W$ $R^2 > 0.85$ use random row-level splits which test interpolation within wells rather than generalisation across wells (Wood, 2021; Al-Bulushi et al., 2009).

The MAPE values for $V_{CL}$ and PORO are astronomically large ($10^{15}$ order of magnitude) due to the presence of depth samples with $V_{CL}$ or PORO values very close to zero in the test wells. Division by near-zero values in the MAPE formula produces numerical overflow. This is a well-documented limitation of MAPE for targets with zero-containing distributions (Miah et al., 2024) and does not reflect any issue with the predictions themselves. MAE and RMSE are the appropriate error metrics for these targets.

# CONCLUSION AND PERSPECTIVES

This dissertation investigated the application of seven supervised machine learning regression algorithms, Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR), Bagging, Extremely Randomised Trees (ExtraTree), and Extreme Gradient Boosting (XGBoost), for the prediction of three key petrophysical parameters from wireline well log data in the Triassic Argilo-Gréseux Inférieur (TAGI) reservoir of the Berkine Basin, eastern Algeria. The three target parameters, water saturation (SW), clay volume (VCL), and effective porosity (PORO), govern hydrocarbon volume estimation, net pay identification, and reservoir simulation input, making their accurate characterisation a central requirement of field development planning. The study was conducted on a consolidated dataset of 12 wells comprising 284,341 depth samples and employed a well-based blind evaluation protocol, in which four test wells were entirely withheld from all training, preprocessing, and hyperparameter tuning decisions. This protocol, which is more rigorous than the random row-level data splits used in the majority of published petrophysical AI studies, ensures that the reported performance metrics reflect genuine inter-well generalisation rather than intra-well interpolation.

The complete methodology followed a sequential pipeline of seven stages: raw data consolidation and null sentinel removal, depth alignment and TAGI interval selection, outlier winsorisation, $\log_{10}$ transformation of resistivity logs, median imputation of missing feature values (fitted on training data only), well-based train/test split (8 training wells, 4 blind test wells), and 5-fold cross-validation for hyperparameter selection. Feature selection retained eight input logs on the basis of a 50% missing-value threshold: GR, CALI, DTC, and the five-depth resistivity array (RT10–RT90). The classical porosity tools RHOB and NPHI were excluded because they exceed the 50% threshold, a data availability constraint that represents the primary limitation of this study. Seven algorithms were implemented in Python 3.10 on the Kaggle cloud platform and evaluated independently for each target, producing 21 model–target combinations assessed by four metrics: $R^2$, RMSE, MAE, and MAPE (the latter being unsuitable for VCL and PORO due to numerical instability near zero-valued samples).

The principal results confirm that gradient-boosted ensemble methods consistently outperform all other algorithms in blind-well generalisation across all three targets. For VCL, GBR achieves the best overall blind-well performance with a test $R^2$ of 0.790 and a mean absolute error of 10.1 percentage points, with all four test wells reaching $R^2 \geq 0.598$ and Well 10 achieving an exceptional $R^2$ of 0.857. This robust and consistent VCL performance is attributed to the near-universal linear relationship between the gamma ray log and clay volume, which the training wells represent faithfully across the basin. Even the linear MLR achieves $R^2 = 0.737$ for VCL, confirming the predominantly linear nature of the GR–VCL relationship in the TAGI and validating the physical basis of the Larionov (1969) empirical correction. These results indicate that VCL prediction is mature enough for immediate operational application in the Berkine Basin, and the predicted VCL curves can be used directly for net-pay estimation and reservoir zonation in wells lacking conventional petrophysical interpretation.

The $S_W$ prediction results are more nuanced and require careful interpretation. The global blind-well test $R^2$ of 0.212 for XGBoost, and 0.320 for the best-performing GBR, are depressed by a statistical anomaly in the Well 9 test well, where the actual SW is near-constant at 0.998 (effectively SW = 1.0 throughout the entire logged TAGI interval). When the target variable has near-zero variance, the $R^2$ denominator approaches zero and any prediction error is amplified by several orders of magnitude, producing the anomalous value of $R^2 = -336$ that dominates the global metric. This is not a model failure but a mathematically degenerate evaluation case that would affect any prediction model regardless of its algorithm. Excluding Well 9, the corrected mean SW $R^2$ across the three remaining test wells is 0.455, with individual well $R^2$ values of 0.369 (Well 10), 0.553 (Well 11), and 0.442 (Well 12), which are consistent with published blind-well benchmarks for this type of heterogeneous clastic reservoir. The near-zero individual Spearman correlations between resistivity logs and SW ($\rho \approx 0.00–0.02$) confirm the low-resistivity pay phenomenon documented by Doghmane et al. (2024) in Algerian Saharan fields, where electrically conductive clay minerals suppress the resistivity contrast between hydrocarbon-bearing and water-saturated intervals, limiting the discriminatory power of resistivity-based approaches. The Random Forest feature importance analysis reveals that GR (39%) and the combined resistivity suite (33%) are the dominant SW predictors, with the resistivity contribution emerging through non-linear ensemble interactions rather than bivariate correlation, a result that could not be captured by deterministic Archie-based models.

Effective porosity prediction achieves moderate but scientifically significant performance, with XGBoost reaching a blind-well test $R^2$ of 0.493 and a mean absolute error of 1.8 porosity percentage points. The 0.245-point $R^2$ improvement from MLR (0.248) to XGBoost (0.493) is the largest algorithm-driven performance gain across all three targets, indicating that non-linear interactions between GR, DTC, CALI, and the resistivity array carry substantial PORO information accessible only to non-linear ensemble methods. At the individual well level, Well 10 achieves $R^2 = 0.731$ and Well 12 achieves $R^2 = 0.582$, confirming that the XGBoost model generalises well to wells with PORO distributions representative of the training cluster. The negative $R^2$ values for Well 9 (−0.366) and Well 11 (−0.237) reflect systematic overestimation of porosity in tight, cemented formations where actual PORO < 0.015 and the DTC-based porosity signal is insufficient to discriminate from moderately porous intervals. This finding directly quantifies the cost of the absent RHOB and NPHI logs: based on comparable published studies that include these tools, the expected PORO $R^2$ improvement ranges from 0.15 to 0.25 points, which would elevate the overall blind-well PORO prediction from moderate to reliable performance.

Three study limitations must be acknowledged. First and most critically, the absence of bulk density (RHOB) and neutron porosity (NPHI) logs from the training feature set, due to their exceeding the 50% missing-value threshold, constrains porosity and water saturation prediction accuracy to below the performance achievable with a complete standard log suite. Second, the study is conducted within a single petroleum province (Berkine Basin, TAGI reservoir); the trained models' direct applicability to other basins and formations require retraining or transfer learning validation. Third, the 5-fold cross-validation used for hyperparameter selection does not fully replicate blind-well generalisation conditions, because training and validation folds contain adjacent depth samples from the same wells. A leave-one-well-out hyperparameter optimisation

strategy would more closely align tuning and evaluation objectives but at substantially higher computational cost. On the basis of these limitations and the results obtained, five research directions are recommended for future work: (i) systematic acquisition of RHOB and NPHI in all future Berkine Basin wells, which represents the highest-value single improvement to prediction accuracy; (ii) leave-one-well-out cross-validation for hyperparameter tuning to better align the optimisation objective with blind-well generalisation; (iii) investigation of sequential deep learning architectures (LSTM, 1D-CNN) that exploit the depth-ordering of log data, which tree-based methods currently ignore; (iv) development of uncertainty quantification outputs (quantile regression, conformal prediction intervals) to support probabilistic reservoir decision-making; and (v) multi-basin transfer learning to extend the validated TAGI methodology to analogous Triassic sandstone reservoirs in adjacent Algerian and Libyan basins.

This dissertation demonstrates that machine learning, when applied with a physically motivated feature selection strategy, a rigorous well-based evaluation protocol, and a systematic comparative framework, provides a valuable and reproducible complement to conventional deterministic petrophysical interpretation in the TAGI reservoir of the Berkine Basin. The primary scientific contribution of this work is the quantified, well-resolved demonstration of blind-well generalisation performance across seven algorithms and three targets, a benchmark that is directly comparable to international standards in the petrophysical AI literature and that provides an evidence-based foundation for the operational adoption of data-driven petrophysical prediction in Algerian reservoir characterisation practice. The complete Python codebase and preprocessed dataset have been structured for full reproducibility, with the expectation that future researchers can extend this framework as additional log data, new wells, and improved algorithms become available.

# REFERENCES


Al-Bulushi, N., King, P. R., Blunt, M. J., & Kraaijveld, M. (2009). Development of artificial neural network models for predicting water saturation and fluid distribution. *Journal of Petroleum Science and Engineering, 68(3–4), 197–208.* DOI: https://doi.org/10.1016/j.petrol.2009.06.017

Al-Mudhafar, W., Rao, D. N., & Srinivasan, S. (2025). Machine learning with hyperparameter optimization applied in facies-supported permeability modeling in carbonate oil reservoirs. *Scientific Reports, 15(1)*. DOI: https://doi.org/10.1038/s41598-025-95490-0

Archie, G. E. (1942). The electrical resistivity log as an aid in determining some reservoir characteristics. *Transactions of the AIME, 146(1), 54–62*.

Boote, D. R. D., Clark-Lowes, D. D., & Traut, M. W. (1998). Palaeozoic petroleum systems of North Africa. In D. S. Macgregor, R. T. J. Moody, & D. D. Clark-Lowes (Eds.), Petroleum Geology of North Africa (Special Publication No. 132, pp. 7–68). Geological Society. DOI: https://doi.org/10.1144/GSL.SP.1998.132.01.02

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24(2), 123–140*. DOI: https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45(1), 5–32*. DOI: https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (pp. 785–794). ACM. DOI: https://doi.org/10.1145/2939672.2939785

Doghmane, M. Z., Belhadj, H., & Ouali, S. (2024). A new approach for estimating water saturation in low-resistivity hydrocarbon-bearing reservoirs using artificial neural networks. *Neural Computing and Applications*. DOI: https://doi.org/10.1007/s00521-024-10777-z

Ellis, D. V., & Singer, J. M. (2007). Well Logging for Earth Scientists *(2nd ed.)*. Springer. DOI: https://doi.org/10.1007/978-1-4020-4602-5

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29(5), 1189–1232*. DOI: https://doi.org/10.1214/aos/1013203451

Fu, L., El-Sonnah, S., & Tian, Z. (2024). Well-log-based reservoir property estimation with machine learning: A contest summary. *Petrophysics – The SPWLA Journal, 65(1), 1–18*. DOI: https://doi.org/10.30632/PJV65N1-2024a6

Galeazzi, S., Point, O., Haddadi, N., Mather, J., & Druesne, D. (2009). Regional geology and petroleum systems of the Illizi-Berkine area of the Algerian Saharan Platform: An overview. *Marine and Petroleum Geology, 26(5), 659–680*. DOI: https://doi.org/10.1016/j.marpetgeo.2008.10.002

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63(1), 3–42*. DOI: https://doi.org/10.1007/s10994-006-6226-1

Haroun Samir, B., Chaouki, S., Hamid, B., & Ali, B. (2024). Petrophysical evaluation based on three-stream convolutional neural networks in the Berkine Basin, Algeria. *Journal of Engineering Research, 12(2), 145–157*. DOI: https://doi.org/10.1016/j.jer.2024.05.031

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning *(2nd ed.)*. Springer. DOI: https://doi.org/10.1007/978-0-387-84858-7

Kannaiah, P. V. D., Ordin, A. A., Andrey, V., Trusov, A., & Yerlan, S. (2023). Machine learning approaches for formation matrix volume prediction from well logs: Insights and lessons learned. *Geoenergy Science and Engineering, 230*. Article 212236. DOI: https://doi.org/10.1016/j.geoen.2023.212086

Larionov, V. V. (1969). Radiometry of Boreholes. NEDRA.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2, 56–67*. DOI: https://doi.org/10.1038/s42256-019-0138-9

Miah, M. I., Ahmed, S., & Rahman, S. (2024). Improved model development and feature ranking for rock permeability prediction by coupling petrophysical log data and ensemble machine learning techniques. *Earth Science Informatics, 17, 3025–3044*. DOI: https://doi.org/10.1007/s12145-024-01593-z

Okon, A. N., Adewole, S. E., & Uguma, E. M. (2020). Artificial neural network model for reservoir petrophysical properties: porosity, permeability, and water saturation prediction. *Modeling Earth Systems and Environment, 7(1), 2373–2390*. DOI: https://doi.org/10.1007/s40808-020-01012-4

Otchere, D. A., Ganat, T. O. A., Gholami, R., & Lawal, M. (2021). A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. *Journal of Natural Gas Science and Engineering, 91*. Article 103962. DOI: https://doi.org/10.1016/j.jngse.2021.103962

Rossi, C., Kalin, O., Clavell, E., & Permanyer, A. (2002). Diagenesis, provenance, and reservoir quality of Triassic TAGI sandstones from the Ourhoud field, Berkine (Ghadames) Basin, Algeria. *Marine and Petroleum Geology, 19(2), 117–142*. DOI: https://doi.org/10.1016/S0264-8172(02)00004-1

Turner, P., Pilling, D., Walker, D., Exton, J., Binnie, J., & Sabaou, N. (2001). Sequence stratigraphy and sedimentology of the late Triassic TAG-I (Blocks 401/402, Berkine Basin, Algeria). *Marine and Petroleum Geology, 18(8), 959–981*. DOI: https://doi.org/10.1016/S0264-8172(01)00039-3

Radwan, A. E., Wood, D. A., & Radwan, A. A. (2022). Machine learning and data-driven prediction of pore pressure from geophysical logs: A case study for the Mangahewa gas field, New Zealand. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(6), 1799-1809. https://doi.org/10.1016/j.petsci.2025.03.023

Yahi, N., Schaefer, R. G., & Littke, R. (2001). Petroleum generation and accumulation in the Berkine Basin, eastern Algeria. *AAPG Bulletin, 85(8), 1439–1467*. DOI: https://doi.org/10.1306/8626CAD7-173B-11D7-8645000102C1865D

Zhao, X., Chen, X., Huang, L., & Su, R. (2022). Logging-data-driven permeability prediction in low-permeable sandstones based on machine learning with pattern visualization: A case study in Wenchang A Sag, Pearl River Mouth Basin. *Journal of Petroleum Science and Engineering, 208*. Article 109609. DOI: https://doi.org/10.1016/j.petrol.2022.110517