Kasdi Merbah University – Ouargla
Faculty of Hydrocarbons, Renewable Energies and Earth and Universe Sciences
Department of Earth and Universe Sciences

Academic Master's Thesis

Domain: Earth and Universe Sciences

Sector: Geology

Specialty: Petroleum Geology

THEME

Real-Time Prediction of Formation Permeability Using Drilling Data and Explainable Machine Learning: A Case Study from the Hassi Tarfa Field

Presented by: Kelai Amel

Publicly defended on: 17/06/2025
In front of the jury:

President: Belaksier M. Salah	**M.C.A**	**Univ. Ouargla**
Examiner: Nmer Zoubida	**M.C.A**	**Univ. Ouargla**
Supervisor: AMEUR-ZAIMECHE Ouafi	**M.C.A**	**Univ. Ouargla**
Co- Supervisor: BOUTHAGHANE Ayoub	**PhD student**	**Univ. Ouargla**

Academic year: 2024/2025

Acknowledgment

With deep humility, I begin by thanking God, whose unwavering grace and strength have guided me through every stage of this journey.

I am profoundly grateful to my supervisors, Dr. Ameur-Zaimeche Ouafi and Mr. Bouthaghane Ayoub, whose insightful guidance, encouragement, and unwavering support were vital to the development of this thesis. Their mentorship has been both a privilege and a source of inspiration.

I would also like to express my sincere appreciation to the jury members for their valuable time, constructive feedback, and for honoring this work with their review.

My thanks extend to the faculty and staff of the Department of Earth and Universe Sciences, whose teaching and dedication have laid the foundation for my academic growth.

Finally, I wish to acknowledge all those—friends, colleagues, and family—whose support, whether seen or unseen, helped bring this work to completion . Your presence, encouragement, and belief in me have made a lasting difference.

dedication

To my parents — the steady rhythm of my heart and the quiet strength behind every step I've taken.
To the ones whose presence brought comfort in moments of exhaustion, and whose prayers lit my path in the darkest times.
This work is a small token of love and gratitude for all you've given me — far beyond words or measure.

And to my siblings,
Your silent support and constant encouragement have meant more than you know.
This thesis is as much yours as it is mine.

To my family — my foundation, my pride —
I dedicate this work, not because it repays your love, but because it could not have been completed without you.

Summary pg

- Acknowledgment
- Dedication
- Abstract
- Résumé
- الملخص
- List of Figures
- List of Tables
- List of Abbreviations
- General Introduction …… I

Chapter 1: Overview of the Study Area

- Introduction 04

1. Geological Setting of the Hassi Tarfa Field 04

1.1 Geographic Location 04

1.2 Geological Framework 04

1.3 Stratigraphy of the Hassi Tarfa Field 05

1.3.1 Paleozoic Sequence 06

1.3.2 Mesozoic Sequence 06

1.3.3 Cenozoic Cover 06

2. Lithological Overview of the Hassi Tarfa Field 07

2.1 Paleozoic 07

2.2 Mesozoic 07

2.3 Cenozoic 08

3. Petroleum System of the Hassi Tarfa Field 09

3.1 Source Rocks 09

3.2 Maturation and Migration 09

3.3 Reservoir Rocks 09

3.4 Covering Rocks 09

3.5 Traps 09

3.6 Hydrocarbon Distribution 10

3.7 Geochemical Characteristics 10

Chapter 2: AI and Machine Learning Concept Generation

- Introduction 12

1. Artificial Intelligence 12

1.1 Subfields of AI 13

2. Machine Learning 13
2.1 Historical Development 14
2.2 Applications of ML 14
2.3 Components of the Learning Process 15
2.4 Classes of Machine Learning 16
2.4.1 Supervised Learning 17
2.4.1.1 Classification Analysis 17
- Core Classification Algorithm 18
2.4.1.2 Regression Analysis 20
- Core Regression Algorithms 21
2.4.2 Unsupervised Learning 23
2.4.2.1 Types of Unsupervised Learning 23
2.4.2.2 Unsupervised Learning Algorithms 24
2.4.3 Semi-Supervised Learning 25
2.4.4 Reinforcement Learning 26
3. Deep Learning 26
3.1 Artificial Neural Network 27
3.1.1 Structure of Neural Network 27
3.2 Algorithms 28
4. Advances and Future Directions: PIML 29
- Conclusion 30

Chapter 3: Review of Real-Time Petrophysical Prediction

- Introduction 32
1. Literature Review – Petrophysical Parameters 33
2. Publication Trends 34
3. Algorithmic Evolution and Usage Trends 36
4. Target Variables in Focus 38
- Conclusion 43

Chapter 4: Materials and Methods

- Introduction 46
1. Conventional Methods 46
1.1 Direct Methods 46
1.1.1 Core Permeability Estimation 46
- 1.1.1.1 Sampling Scales 46

1.1.1.2 Absolute Permeability 47
1.1.1.3 Relative Permeability 47
1.2 Indirect Methods 47
1.2.1 Wireline-Log-Based Permeability Estimation 47
1.2.1.1 Empirical Correlations from Porosity and Water Saturation…… 47
1.2.2 Nuclear Magnetic Logging (NML) 48
1.2.3 Geochemical Logging Tools (GLT) 49
1.2.4 Stoneley Wave Analysis from Sonic Logs 49
1.2.5 Pressure Transient Analysis with Formation Testers (RFT, MDT)…….. 49
1.2.6 Well Testing (Dynamic Field Methods) 50
1.2.6.1 Conventional Testing 51
1.2.6.2 Advanced Testing Techniques 51
2. Unconventional Method 51
2.1 Data Collection 52
2.2 Data Cleaning 52
2.3 Descriptive Statistics 52
2.3.1 Univariate Statistics 52
2.3.2 Bivariate Data 53
2.3.3 Multivariate Data 53
2.4 Data Splitting 53
2.5 Algorithm Selection 54
2.6 Validation Criteria 55
2.7 Model Interpretability Tools 56
2.7.1 SHAP (SHapley Additive Explanations) 56
2.7.2 Feature Importance 56
3. Software Tools 56
3.1 Python 56
3.2 Excel 56

Chapter 5: Results & Discussions

- Introduction 59
1. Model Deployment Setup 59
2. Results and Discussion 59
2.1 Univariate Statistics 59
2.2 Bivariate Statistics 60
2.3 Model Performance Evaluation 61

2.3.1 Training Performance 61
2.3.2 Testing Performance 61
2.4 Best Feature Subset Performance 63
2.5 Feature Importance and SHAP Value Analysis 63
Recommendations……. 66
General Conclusion……. 67
Bibliography……..……

Abstract

This study investigates the use of ensemble machine learning models—XGBoost, CatBoost, AdaBoost, and Gradient Boosting—to predict formation permeability from real-time drilling parameters. Using a dataset of 183 samples, CatBoost showed the best performance (R^2 = 0.837). After feature selection, XGBoost achieved the highest accuracy (R^2 = 0.9655). SHAP analysis identified Torque, Flow Rate, and Rate of Penetration as key predictors. The results highlight the benefits of feature selection and explainable AI for real-time reservoir characterization.

Keywords: Formation permeability, machine learning, XGBoost, CatBoost, drilling parameters, SHAP, feature selection, reservoir characterization.

Résumé

Cette étude explore l'utilisation de modèles d'apprentissage automatique ensemblistes—XGBoost, CatBoost, AdaBoost et Gradient Boosting—pour prédire la perméabilité de la formation à partir des paramètres de forage en temps réel. En utilisant un jeu de données de 183 échantillons, CatBoost a montré les meilleures performances (R^2 = 0,837). Après sélection de caractéristiques, XGBoost a atteint la plus grande précision (R^2 = 0,9655). L'analyse SHAP a identifié le couple Couple, Débit et Taux de pénétration comme principaux prédicteurs. Les résultats soulignent les avantages de la sélection de caractéristiques et de l'IA explicable pour la caractérisation en temps réel des réservoirs.

Mots-clés : Perméabilité de la formation, apprentissage automatique, XGBoost, CatBoost, paramètres de forage, SHAP, sélection de caractéristiques, caractérisation des réservoirs.

الملخص

تستعرض هذه الدراسة استخدام نماذج التعلم الآلي الجماعي—XGBoost و CatBoost و AdaBoost و Gradient Boosting—لتنبؤ نفاذية التكوين استنادًا إلى معلمات الحفر في الوقت الفعلي. باستخدام مجموعة بيانات تحتوي على 183 عينة، أظهر نموذج CatBoost أفضل أداء (R^2 = 0.837). بعد تطبيق اختيار السمات، حقق نموذج XGBoost أعلى دقة (R^2 = 0.9655). أظهر تحليل SHAP أن العزم، معدل التدفق، ومعدل الاختراق هي المتغيرات الرئيسية المؤثرة. تؤكد النتائج على أهمية اختيار السمات والذكاء الاصطناعي القابل للتفسير في تحسين التصنيف الفعلي للآبار.

الكلمات المفتاحية: نفاذية التكوين، التعلم الآلي، XGBoost، CatBoost، معلمات الحفر، SHAP، اختيار السمات، تصنيف الخزانات.

List of Figures	Pg
• Figure 01. Location of Hassi Tarfa oil field (Aoun et al., 2022)	**04**
• Figure 02. Principal Geological Features of Algeria (SONATRACH, 2005)	**05**
• Figure 03. Lithological column of HTF	**08**
• Figure 04. Petroleum System of Hassi Messaoud and Hassi Tarfa (Sonatrach, Exploration Division, 2007)	**10**
• Figure 05. Map of AI Fields Javaid, Haleem, Khan, & Suman, 2023)	**13**
• Figure 06. Definition of Machine Learning (Mitchell, 1997)	**14**

- Figure 07. The six main application areas of machine learning in oil and gas **15**
- Figure 08. Components Learning Process **16**
- Figure 09. Traditional machine learning types **16**
- Figure 10. Structure of a Random Forest **18**
- Figure 11. Extreme Gradient Boosting Machine (XGBoost) **19**
- Figure 12. SVM Decision Boundary **20**
- Figure 13. The difference between classification and regression methods **20**
- Figure 14. Linear regression and correlation analysis **21**
- Figure 15. Types of Unsupervised Learning **23**
- Figure 16. Example of Clustering **24**
- Figure 17. Autoencoder Components **24**
- Figure 18. Workflow of K-means Clustering **25**
- Figure 19. Machine learning and deep learning performance with data volume **26**
- Figure 20. Biological Neural Network present in Human Body **27**
- Figure 21. A three-layered Artificial Neural Network **28**
- Figure 22. Basic convolutional neural network structure (LeCun, 1989) **29**
- Figure 23. (a) Simple RNN structure with feedback loop; (b) Unfolded RNN across time steps **29**
- Figure 24. Overview of physics-informed machine learning (Hao, Liu, Zhang, Ying, Feng, Su, & Zhu, 2023). **30**
- Figure 25. Temporal Keyword Analysis of Machine Learning Applications in Petrophysical Prediction (2019–2025) **34**
- Figure 26. Temporal distribution of relevant research articles (2019–2025) **36**
- Figure 27. Frequency of algorithms used between 2019–2025 **38**
- Figure 28. Target Petrophysical Parameters Modeled in Real-Time ML Studies **39**
- Figure 29. Charts for estimating permeability from porosity and water saturation **48**
- Figure 30. RFT tool setup and schematic of sampling system **50**
- Figure 31. Typical Impulse test plot **51**
- Figure 32. Methodological framework **52**
- Figure 33. Typical Machine Learning Data Split: Training, Validation, and Testing Sets **54**
- Figure 34. Flowchart summary of Permeability Prediction method **57**
- Figure 35. Pair Plot & Correlation Heatmap of Drilling Parameters **60**

- Figure 36. Training vs. Testing Performance of Regression Models **62**
- Figure 37. Feature Importance Rankings for Permeability Prediction Models **64**
- Figure 38. SHAP Summary Plots **65**

List of Tables **Pg**

- **Table 01.** Comparison of Regression Techniques 22
- **Table 02.** Summary of Research Studies on Real-Time Petrophysical Parameter Estimation Using Artificial Intelligence 39
- **Table 03.** Univariate Analysis 60
- **Table 04.** Initial Training Performance Metrics for Permeability Prediction Models 61
- **Table 05.** Initial Testing Performance Metrics for Permeability Prediction Models 61
- **Table 06.** Best Performance Metrics with Optimized Feature Subsets for Each Model 63

List of Abbreviations

• AdaBoost (AB)
• Artificial Intelligence (AI)
• Artificial Neural Networks (ANN)
• Adaptive Neuro-Fuzzy Inference System (ANFIS)
• CatBoost (CB)
• Convolutional Neural Networks (CNNs)
• Drilling Data (DD)
• Deep Learning (DL)
• Deep Neural Networks (DNN)
• Decision Tree (DT)
• Gradient Boosting Regressors (GBRs)
• Gas-While-Drilling (GWD)
• Hybrid Perceptron Deep Neural Network (HPDNN)
• Least Squares Support Vector Machine (LS-SVM)
• Machine Learning (ML)
• Multiple Linear Regression (MLR)
• Natural Language Processing (NLP)
• Polynomial Regression (PR)
• Particle Swarm Optimization - Neural Network (PSO-NN)
• Random Forest (RF)
• Reinforcement Learning (RL)
• Root Mean Squared Error (RMSE)
• Recurrent Neural Networks (RNNs)
• Rate of Penetration (ROP)
• Support Vector Machine (SVM)
• Total Organic Carbon (TOC)
• Extreme Gradient Boosting (XGBoost)

General Introduction

In the context of escalating global energy demands and the increasingly complex nature of hydrocarbon extraction, the accurate assessment of subsurface reservoirs remains a cornerstone of efficient and sustainable resource development. Central to this assessment are key petrophysical parameters such as porosity, fluid saturation, and particularly permeability, which governs a reservoir's ability to transmit fluids. Among these, permeability plays a decisive role in determining reservoir productivity, guiding drilling strategies, forecasting production, and optimizing recovery processes (Chen & Mohaghegh, 2018).

Traditionally, permeability has been estimated through well log interpretation and laboratory analysis of core samples. Although these methods have provided valuable insights for decades, they are often limited by high operational costs, delayed data availability, and sensitivity to borehole conditions (Helle et al., 2001). In many cases, they offer only static information, which can be insufficient for capturing the spatial and temporal variability of complex reservoirs.

As the oil and gas industry evolves to meet modern challenges, there is a growing shift toward digital and data-driven approaches. Artificial intelligence and machine learning have emerged as powerful tools for modeling subsurface properties, especially permeability, in cases where traditional methods are less effective or too time-consuming. These technologies can uncover hidden patterns and relationships within large datasets, offering improved accuracy and efficiency in petrophysical analysis (Santosh et al., 2020). However, for these models to be effective and trustworthy, they must be grounded in solid geological understanding.

Geology remains the essential foundation of any reservoir evaluation. A proper understanding of the subsurface—its stratigraphy, lithology, structure, and evolution—is fundamental to interpreting petrophysical properties. Geological knowledge ensures that any data-driven predictions align with the physical realities of the reservoir and supports more meaningful, context-aware modeling.

This thesis focuses specifically on the prediction of formation permeability using a data-driven approach applied to conventional datasets from the Hassi Tarfa Field in the Algerian Sahara. It aims to demonstrate how geological understanding, combined with modern analytical techniques, can enhance the accuracy, reliability, and interpretability of permeability estimation.

The thesis is structured into five chapters:

Chapter 1 provides a general geological overview of the study area, highlighting its setting and relevance to reservoir development. This chapter sets the context for both conventional interpretation and data-driven analysis.

Chapter 2 introduces the foundational concepts of artificial intelligence and machine learning, explaining their development and growing role in modeling subsurface properties. It discusses different types of learning and their applications within and beyond the energy sector.

Chapter 3 presents a critical review of existing research related to the use of machine learning in petrophysical prediction, with particular emphasis on permeability estimation. It traces the evolution of methodologies, data types, and evaluation techniques while identifying emerging trends in the field.

Chapter 4 outlines the methodology adopted in this study, including data selection, preprocessing steps, and the modeling process. It details how permeability was predicted from available data, and how the models were evaluated for accuracy and reliability.

Chapter 5 presents and interprets the results of the study, analyzing model performance and discussing the outcomes in light of the geological context. The findings are evaluated for their implications in reservoir characterization and future application potential.

This structure allows the thesis to progress logically from foundational geological and theoretical concepts to applied data analysis and interpretation, offering a comprehensive approach to improving permeability prediction in subsurface evaluation.

Chapter 01

Overview Of The Study Area

Introduction

The Algerian Sahara is one of the most important hydrocarbon provinces in North Africa, characterized by vast sedimentary basins that have accumulated thick sequences of sediment over geological time. This stable cratonic platform, located south of the Atlas Mountains, hosts multiple oil and gas fields that significantly contribute to Algeria's energy production. The geological framework of the region is marked by diverse structural features and a complex stratigraphic history, creating favorable conditions for hydrocarbon generation, migration, and trapping (Makhous et al., 2009).

These sedimentary basins extend across a large area, reflecting a dynamic tectonic and sedimentary evolution from the Paleozoic to the Cenozoic eras. Understanding the geological setting and basin architecture is crucial for effective exploration and exploitation of hydrocarbons. The region remains a focus of ongoing research and development due to its vast resource potential and strategic importance.

Geological Setting of the Hassi Tarfa Field

1.1. Geographic Location

The Hassi Tarfa area is located in the Ouargla Province in southeastern Algeria, approximately 650 kilometers south of Algiers and about 44 kilometers south of Hassi Messaoud. It lies within Block 427, as defined by the Sonatrach concession map. The field is geographically bounded by the parallels 31° and 32° North, and the meridians 6° and 7° East. Hassi Tarfa is positioned in a transitional zone between the Hassi Dzabat permit area and the Hassi Messaoud field.

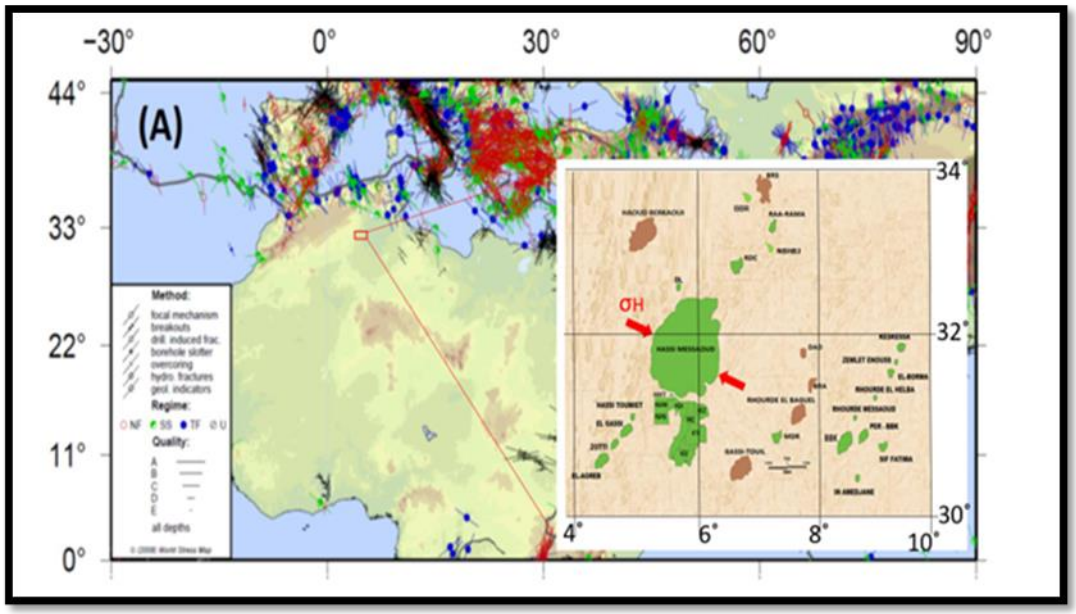


Figure 01. Location of Hassi Tarfa oil field (Aoun et al, 2022).

1.2. Geological Framework

The Hassi Tarfa field is located within the Triassic Province, positioned to the south of the Hassi Messaoud reservoir, along the structural trend that extends from El Gassi to El Agreb and Hassi Messaoud. This region is part of the broader geological context of the southeastern

Algerian sedimentary basin, which is known for its significant hydrocarbon potential. The field is bordered by several key geological features:

- To the north and northeast, it is adjacent to the Hassi Messaoud field, one of Algeria's largest oil fields, which shares similar geological characteristics,
- To the west, the Hassi D'zabat anticlinal structure marks the boundary, playing a significant role in fluid migration and reservoir formation,
- To the east, it is bordered by the Mesdar field, which contributes to the regional petroleum system,
- To the south, the field extends towards the El Gassi field, further linking it to the broader regional trend.

The geological setting of the Hassi Tarfa field, situated within the Triassic Province, is marked by a combination of faulting and folding that provides favorable conditions for hydrocarbon accumulation. The proximity to other major fields, such as Hassi Messaoud, suggests a regional continuity in the subsurface structures, which are integral to the hydrocarbon migration and trapping mechanisms in the area.

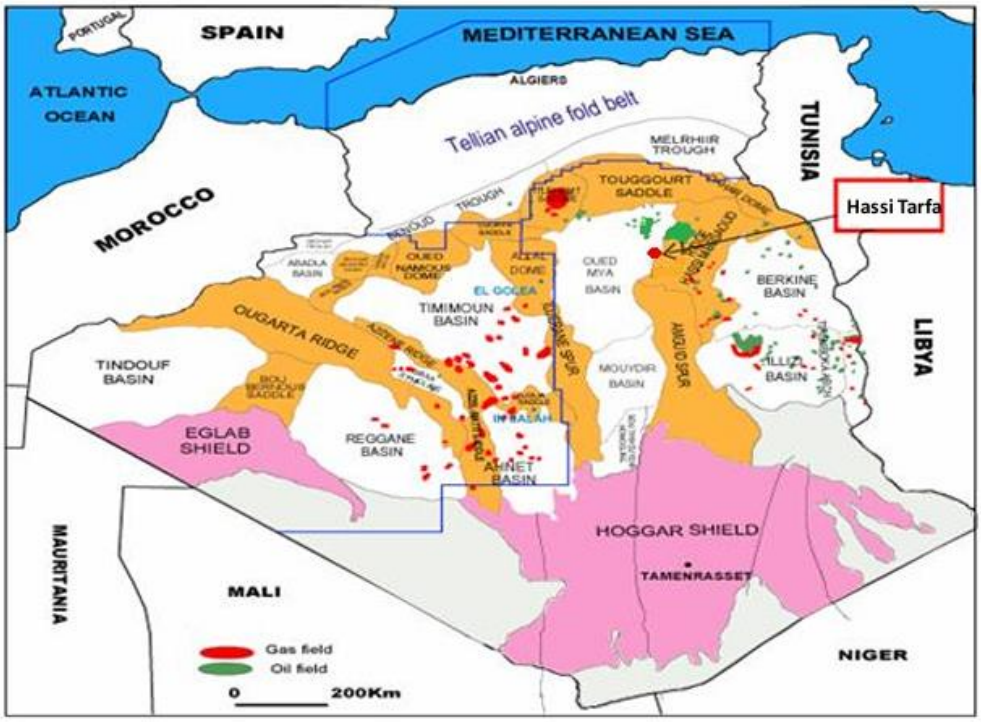


Figure 02. Principal Geological Features of Algeria (SONATRACH, 2005).

1.3. Stratigraphy of the Hassi Tarfa Field

The stratigraphy of the Hassi Tarfa field presents a complex geological history marked by significant depositional events spanning from the Paleozoic to the Cenozoic**.** The region is primarily composed of Mesozoic deposits, which form a substantial portion of the geological sequence, with a total thickness of approximately 3118 meters. These deposits unconformably overlie Paleozoic formations that are about 407 meters thick, while a thin layer of Tertiary detrital deposits (around 300 meters**)** unconformably rests atop the Mesozoic sequence.

1.3.1. Paleozoic Sequence

The Paleozoic strata are relatively thin in Hassi Tarfa, with an average thickness of about 400 meters (Saiad & Derbal, 2016). This interval mainly comprises Cambro-Ordovician detrital formations, dominated by quartzitic sandstones deposited in fluvio-deltaic to shallow marine environments. The source of these sediments was from southern cratonic areas, progradating northward across an eroded Pan-African basement (Boumelit & Lefilef, 2019).

The Cambro-Ordovician period was also punctuated by minor volcanic episodes, although their lateral extent remained limited. The Upper Ordovician reflects glacial influence, with significant erosional surfaces linked to tectonic compression events. The Silurian succession, following a major marine transgression, is marked by organic-rich black shales deposited under anoxic conditions — an important potential source rock for hydrocarbons (Sahraoui & Mahdjoudi, 2024).

1.3.2. Mesozoic Sequence

The Mesozoic deposits form the thickest part of the stratigraphy at Hassi Tarfa, with a cumulative thickness exceeding 3,100 meters. The base of the Mesozoic is characterized by Upper Triassic argillaceous and sandy formations deposited under dominantly continental (fluvial) conditions following a regional angular unconformity (Boumelit & Lefilef, 2019).

Later, evaporitic environments prevailed during the Late Triassic to Early Jurassic, associated with rifting episodes and minor volcanic activities across the Saharan platform. The evaporites comprise dolomites, marls, oolitic limestones, and marno-carbonate intervals, which serve as important regional seals.

The Lower Cretaceous is marked by widespread deposition of quartzitic sandstones due to a major regressive event from the north. These sandstones are key reservoir units, with enhanced secondary porosity due to diagenetic processes. During the Aptian, a brief transgressive event resulted in the deposition of carbonates, indicating temporary stabilization of marine conditions.

The Cenomanian-Turonian sequence signals a renewed transgressive phase, progressing from restricted lagoonal settings to fully open marine conditions, recording important shale and limestone deposition that influenced regional hydrocarbon systems

1.3.3. Cenozoic Cover

The Tertiary succession is relatively thin, averaging around 300 meters. It consists mainly of shallow marine and lagoonal carbonates and clastics, deposited during the Paleocene and Eocene (Saiad & Derbal, 2016).

The Cenozoic period coincides with the effects of the Alpine orogeny, during which compressive stresses reactivated older structures, creating favorable conditions for structural trapping of hydrocarbons in the deeper formations.

2. Lithological Overview of the Hassi Tarfa Field

2.1. Paleozoic

Cambrian:

- R3**:** Base composed of coarse to very coarse white sandstones, locally micro-conglomeratic with clay-rich cement. Minor ferruginous sandstones and silty clays occur.
- R2**:** Medium to coarse, quartzitic sandstones with mica content and silt interbeds. Poor sorting and abundant clay cement present.
- Ra**:** White to very coarse, compact quartzitic sandstones with local pyrite. Sparse gray to black clay seams.
- Ri**:** Fine to medium, pyrite-rich white sandstones with silico-quartzitic texture and micaceous clay streaks. Common horizontal fractures with Tigillites.

Ordovician:

- Alternation Zone**:** Interbedded dark silty-micaceous clays and medium-grain siliceous sandstones, grading into grayish-white siltstones.
- El Gassi Clays**:** Dense, dark gray clays with silty, micaceous texture and occasional glauconitic sandstones or volcanic fragments.
- El Atchane Sandstones**:** Compact fine to medium quartzitic sandstones, locally glauconitic or bituminous, with dark clay interlayers.
- Hamra Quartzites**:** Gray to white quartzites with interbedded micaceous clays and common pyritized fractures and Tigillites.
- Ouargla Sandstones**:** Very fine, silico-quartzitic sandstones, often mottled with pyrite, underlain by volcanic rocks and silty clays.

2.2. Mesozoic

Triassic:

- Volcanic Series**:** Basaltic and andesitic flows interbedded with reddish, clayey siltstones.
- Clayey Unit**:** Variegated red to brown silty clays, interspersed with dolomitic and anhydritic beds.
- Saliferous Unit**:** Thick halite layers alternating with anhydrite and dolomitic clays.

Jurassic:

- **Lias:**
  - Horizon B**:** Gray to greenish-gray microcrystalline limestone with interbedded clays and a basal anhydrite layer.
  - LS2**:** White halite with reddish plastic clays.
  - LD2**:** Alternating dolomite, variegated clay, and white-gray anhydrite.
  - LS1: White salt alternating with anhydrite and clays, topped by massive anhydrite.
  - LD1**:** Sequences of clay, anhydrite, and dolomite.

Dogger:

- Lagoonal Facies**:** White anhydrite, red clays, and dolomitic limestone.
- Clayey Facies**:** Indurated, brown silty clays with carbonate content, anhydrite, and fine sandstone layers.

- **Malm:** Brown to greenish silty clays with fine sandstones, dolomite, anhydrite, and rare lignite.

2.3. Cenozoic

- Eocene-Tertiary**:** Sequence of clayey-sandy, carbonate, or dolomitic deposits, with red to brown coloration, sometimes lagoonal. Occasional presence of gypsum or marl layers.

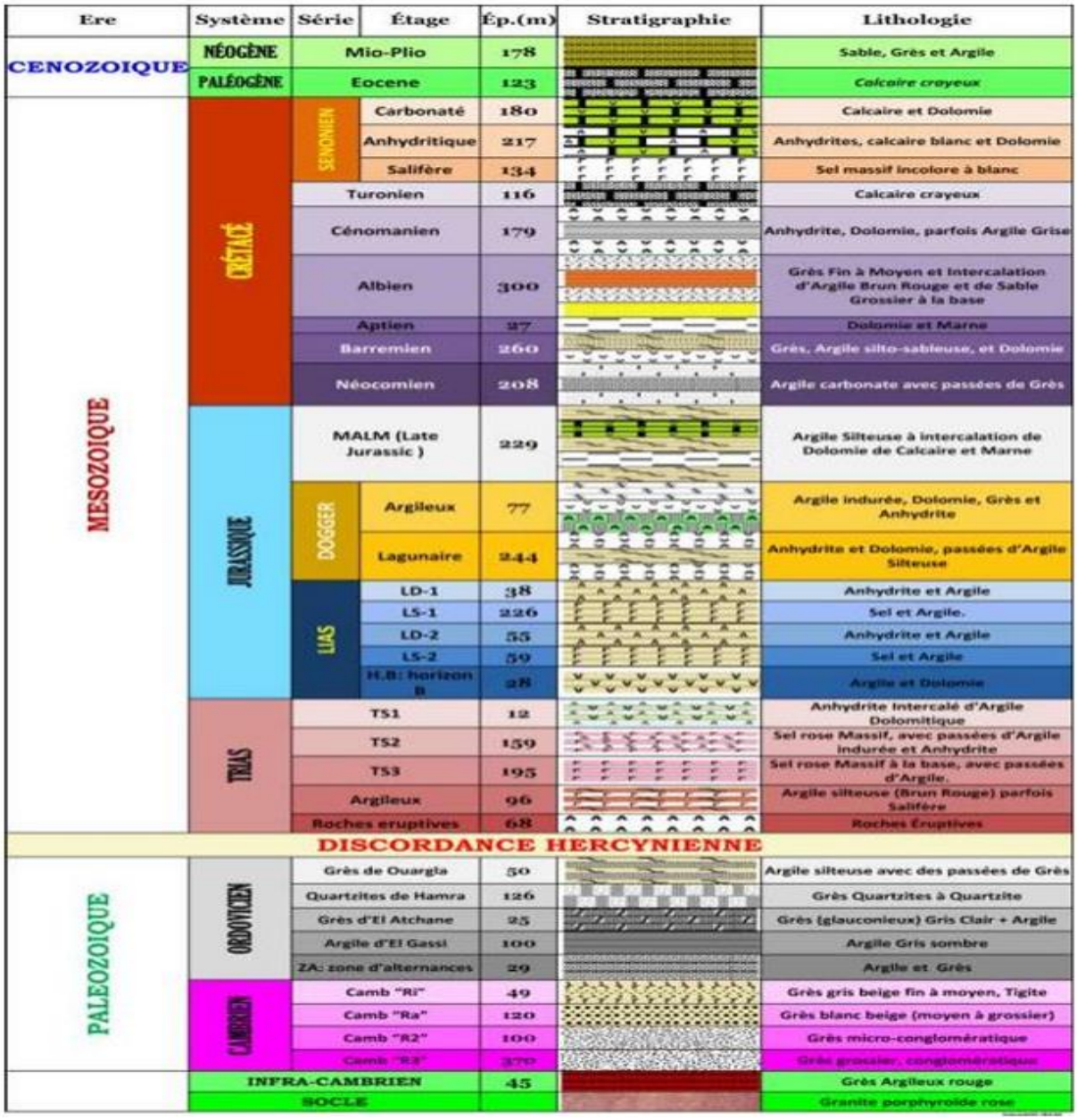


Figure 03. Lithostratigraphic column

3.Petroleum System of Hassi Terfa Field

3.1 Source Rocks

The primary source rocks in the Hassi Terfa field are the Silurian organic-rich shales. These shales are responsible for 80-90% of the Paleozoic hydrocarbon sources in North Africa. The radioactive "Hot Shales" at the base of the Silurian, dated from Rhuddanian to Aéronien, were deposited across a large portion of the ancient northern Gondwanan margin. These shales are highly rich in Total Organic Carbon (TOC), with concentrations reaching up to 17%. The Silurian source rock is characterized by kerogen type II, a mix of algae, amorphous organic matter, and some Chitinozoans. The Silurian remains the primary source rock due to its high organic richness, with a TOC value of 14%. Additional potential source rocks include the Azzel and El-Gassi shales, though they are less significant.

3.2 Maturation and Migration

Hydrocarbon maturation and migration predominantly took place between the Late Jurassic and Early Cretaceous periods. The relatively shallow burial of the Silurian shales during the Paleozoic era contributed to the preservation of their hydrocarbon potential. Migration pathways were largely controlled by fault systems and erosion surfaces, enabling the movement of hydrocarbons into the Cambrian-Ordovician reservoir traps.

3.3 Reservoir Rocks

The principal reservoir in the Hassi Terfa field is the Ordovician Hamra Quartzites. Despite their compact nature and generally poor petrophysical properties, these formations hold substantial quantities of oil and gas. Additional reservoirs are found within the Ra and Ri lithozones of the Cambrian succession, which also contribute to the hydrocarbon accumulation in the field.

3.4 Covering Rocks

The Jurassic clayey-saline formations (Lias) and Triassic volcanic rocks act as effective seals for the Hamra Quartzites reservoir. Additionally, the El-Gassi shales provide a competent sealing barrier for the Cambrian reservoirs, ensuring effective hydrocarbon entrapment and preservation.

3.5 Traps

Hydrocarbons in the Hassi Terfa field are primarily accumulated in **structural traps**, particularly anticlines, which formed due to tectonic movements. These traps feature lower pressure and temperature conditions than the source rocks, promoting hydrocarbon accumulation**.**

3.6 Hydrocarbon Distributin

The hydrocarbons in the field consist primarily of light oil**,** which is of marine origin and was deposited under reducing conditions. In the northeastern and western parts of the basin, the oil remains in liquid form, while in the southwestern part**,** it transitions into dry gas.

3.7 Geochemical Characteristics

Geochemical analyses show that the oil is marine-derived and was deposited under reducing conditions. The migration and accumulation of hydrocarbons were largely driven by tectonic activity and subsidence during the Mesozoic**.**

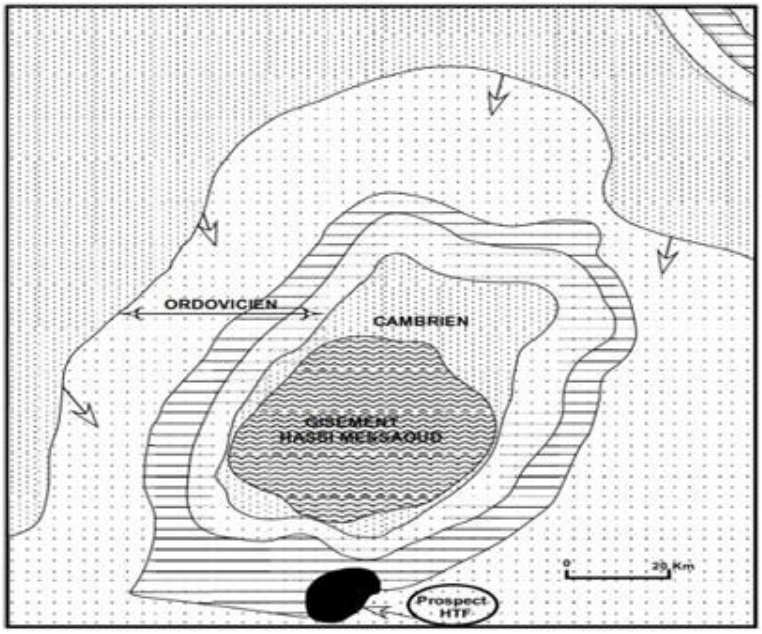


Figure04. Petroleum System of Hassi Messaoud and Hassi Tarfa (Sonatrach, Exploration Division, 2007).

Chapter 02

AI and Machine Learning Concept Generation

Introduction

The oil and gas industry faces increasing challenges, including resource depletion, rising operational costs, and environmental concerns. As conventional methods struggle to keep pace with these complexities, Machine Learning (ML) and Artificial Intelligence (AI) have emerged as transformative tools to enhance efficiency, reduce risks, and optimize decision-making (Hyne, 2012).

ML models can analyze vast amounts of data, enabling predictive maintenance, real-time drilling optimization, and reservoir characterization. These applications improve operational reliability, minimize failures, and reduce downtime, offering a cost-effective alternative to traditional empirical approaches (Meadows et al., 2004). Furthermore, AI-driven automation enhances safety and reduces human intervention in high-risk environments.

With the industry's growing emphasis on digital transformation, integrating ML into oil and gas operations is no longer optional but essential for sustainability and competitiveness (Ameur-Zaimeche et al., 2022). By leveraging advanced data analytics, companies can maximize production efficiency while minimizing environmental impact, aligning with the global push toward more responsible energy management (Meadows et al., 1972).

1. Artificial Intelligence

Artificial Intelligence (AI) is a multidisciplinary field dedicated to developing systems capable of performing cognitive functions typically associated with human intelligence, such as reasoning, learning, and problem-solving (Nilsson, 2009). AI research has evolved significantly, with foundational work documented by Nilsson (2009) and later advancements in machine learning explained by Domingos (2015) and Mitchell (2019). The global race for AI dominance has been extensively analyzed, particularly in the context of economic and technological leadership (Lee, 2018). Meanwhile, Ford (2018) highlights perspectives from leading AI researchers, offering insight into both theoretical advancements and practical applications.

AI research is supported by major professional societies, including the Association for the Advancement of Artificial Intelligence (AAAI) and the ACM Special Interest Group in Artificial Intelligence (SIGAI), while ethical concerns surrounding AI's societal impact are addressed by organizations such as the Partnership on AI. Key academic contributions appear in conferences like the International Joint Conference on Artificial Intelligence (IJCAI) and journals such as Artificial Intelligence and IEEE Transactions on Pattern Analysis and Machine Intelligence, ensuring continuous advancements in the field.

As AI has evolved, it has branched into several specialized subfields that focus on different aspects of intelligence, such as perception, learning, and decision-making. Among these, Machine Learning (ML) has emerged as one of the most transformative areas, driving advancements in automation, predictive analytics, and autonomous systems. However, before diving into ML in detail, it is essential to understand the broader landscape of AI and its key subfields.

1.1 Subfields of Artificial Intelligence

Artificial Intelligence is a broad field encompassing various subdomains, each contributing to different aspects of intelligent behavior. Some of the major subfields include:

- **Machine Learning (ML):** A core subfield of AI, ML enables computers to improve their performance by learning from data without being explicitly programmed. It allows systems to recognize patterns, make predictions, and adapt over time based on experience.
- **Computer Vision:** This field focuses on enabling machines to interpret and understand visual information from images and videos. Applications include facial recognition, object detection, and autonomous vehicles.
- **Natural Language Processing (NLP):** NLP allows machines to understand, process, and generate human language. It is used in applications such as speech recognition, machine translation, and sentiment analysis.
- **Deep Learning:** A subset of ML, deep learning utilizes artificial neural networks to model complex relationships in data. It is particularly effective in image recognition, natural language understanding, and autonomous decision-making.
- **Reinforcement Learning:** In this type of AI, an agent learns by interacting with its environment and receiving rewards or penalties based on its actions. This approach is widely used in robotics, game playing, and autonomous systems.

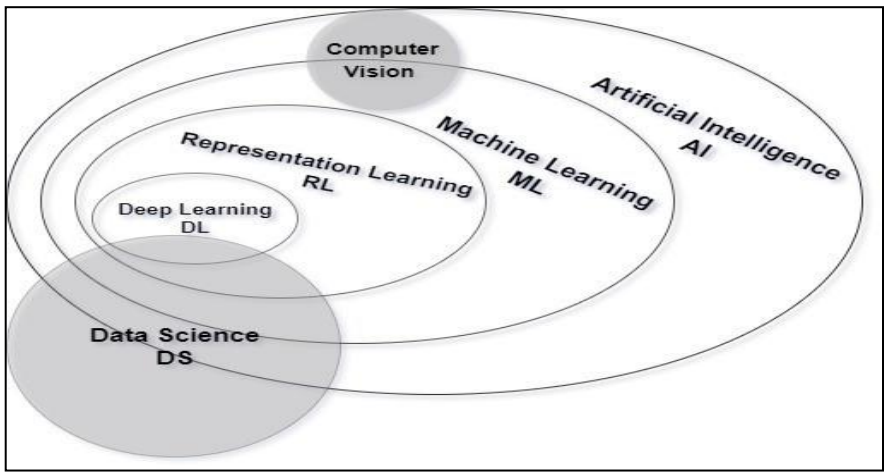


Figure 05. Map of AI Fields (Javaid, Haleem, Khan, & Suman, 2023)

2. Machine learning

A subfield of artificial intelligence that focuses on building models by learning directly from empirical data, rather than relying on manually crafted rules or explicit programming. It uncovers patterns and relationships among variables using algorithms such as neural networks, decision trees, and clustering techniques. These models are trained on datasets to extract insights and generate predictions, making them especially useful for real-time applications.

Formally, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 1997).

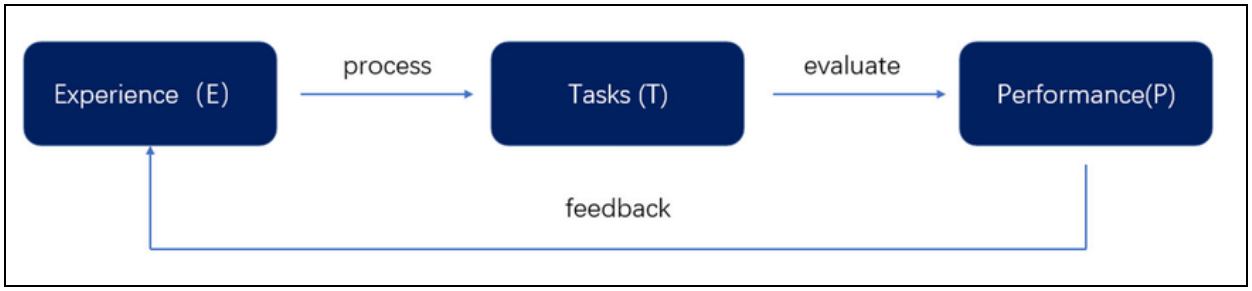


Figure 06. Definition of Machine Learning (Mitchell, 1997)

2 .1 Historical Development of Machine Learning

The term Machine Learning was coined by Arthur Samuel in 1952. In 1957, Frank Rosenblatt developed the *perceptron* at the Cornell Aeronautical Laboratory, building on the work of Donald Hebb and Arthur Samuel. Initially designed as a hardware-based system rather than a software program, it was implemented on the Mark 1 perceptron, a custom-built computer developed for the IBM 704, enabling the transfer of machine learning algorithms across different machines (Foote, 2019).

The foundation of pattern recognition emerged in 1967 with the nearest neighbor algorithm, which provided early solutions for route optimization problems, such as those faced by traveling salespeople. However, progress in machine learning remained limited through the 1970s, largely due to the dominance of Von Neumann architecture, which prioritized traditional computing paradigms over neural networks.

A breakthrough came in 1982 when John Hopfield introduced a network of bidirectional connections, closely mimicking biological neurons—an approach still fundamental in deep learning. That same year, Japan's commitment to advancing neural networks spurred increased research and funding in the United States (Foote, 2019).

In the mid-to-late 1980s, significant advancements were made, including *NETtalk* (1985) by Terrence Sejnowski, which learned to pronounce written English text, and the introduction of backpropagation *(*1986), which greatly enhanced neural network training. In 1989, Yann LeCun integrated backpropagation into convolutional neural networks (CNNs*)* for handwriting and optical character recognition (OCR).

A major milestone occurred in 1997 when IBM's Deep Blue became the first machine to defeat a reigning world chess champion under standard time constraints, demonstrating the potential of machine learning in surpassing human cognitive abilities.

Entering the 21st century, businesses increasingly recognized the potential of machine learning, leading to extensive investments in research and development. As a result, machine learning has evolved into a critical field, driving innovations across industries and accelerating artificial intelligence advancements.

2.2 Applications of Machine Learning

Machine learning (ML) has become a transformative force across various industries, driving automation, optimizing complex processes, and enhancing predictive capabilities. By

leveraging vast datasets and advanced algorithms, ML enables more accurate decision-making and operational efficiency. In sectors where large volumes of data are generated, such as healthcare, finance, manufacturing, transportation, and energy, ML applications are proving indispensable.

- **Healthcare** – ML assists in medical diagnosis by detecting diseases such as cancer and heart conditions, accelerates drug discovery, and enables personalized treatment plans.
- **Finance** – It is used for fraud detection, algorithmic trading, and credit risk assessment to enhance security and financial decision-making.
- **Manufacturing** – ML improves quality control by identifying defects, optimizes supply chains, and enhances process automation for increased efficiency.
- **Transportation** – It powers autonomous vehicles, optimizes delivery routes, and enables predictive maintenance for fleet management and logistics.
- **Oil and Gas Industry** – In energy exploration and production, ML is revolutionizing operations by improving subsurface characterization, optimizing drilling processes, enhancing production forecasts, and enabling predictive maintenance to prevent equipment failures. These advancements increase efficiency, reduce risks, and support data-driven decision-making. The six key application areas of ML in the oil and gas sector are detailed in Figure (07)

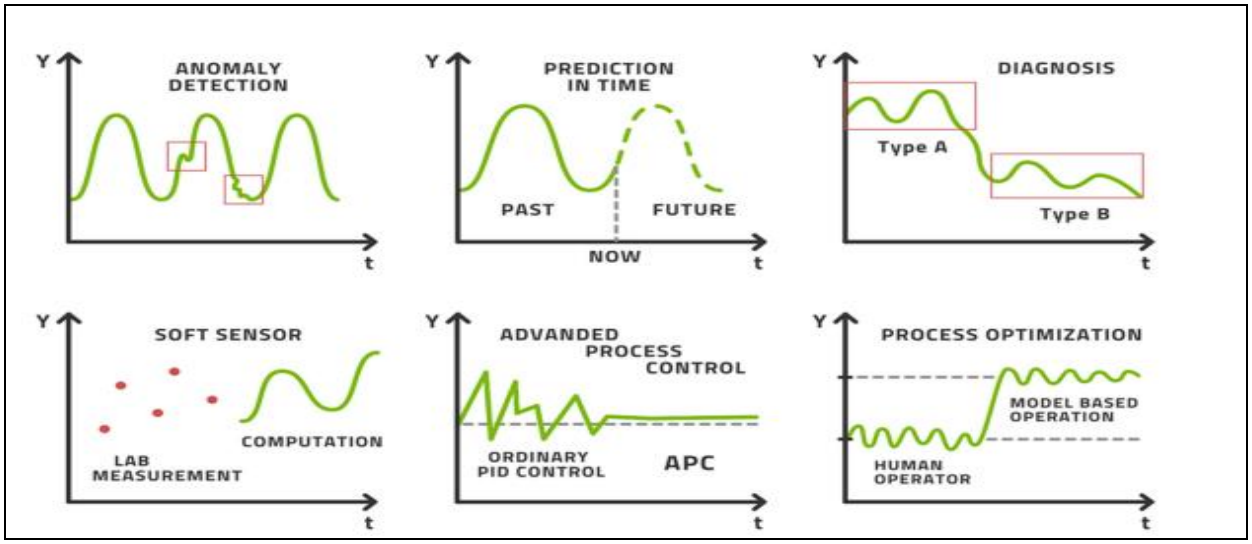


Figure 07. The six main application areas of machine learning in oil and gas

2.3 Components of the Learning Process

The learning process, whether in humans or machines, consists of four fundamental components: data storage, abstraction, generalization, and evaluation.

- **Data Storage** – Effective learning requires the ability to store and retrieve large amounts of data. In humans, this function is carried out by the brain through electrochemical signals, while computers use storage devices such as hard drives, flash memory, and RAM to access and process information efficiently.
- **Abstraction** – This involves extracting useful knowledge from stored data by applying existing models or creating new ones. In machine learning, this process is known as

training, where a model is fitted to a dataset to transform raw data into a meaningful, summarized form that can be used for predictions or decision-making.

- **Generalization** – Once a model has been trained, it must be capable of applying learned patterns to new, unseen data. Generalization ensures that the system can perform well on similar but not identical tasks. The goal is to identify key properties of the data that remain relevant across different scenarios.
- **Evaluation** – The effectiveness of the learning process is measured through evaluation, where the model's performance is tested using various metrics. This feedback loop is essential for refining the model, improving accuracy, and optimizing its application to real-world tasks.

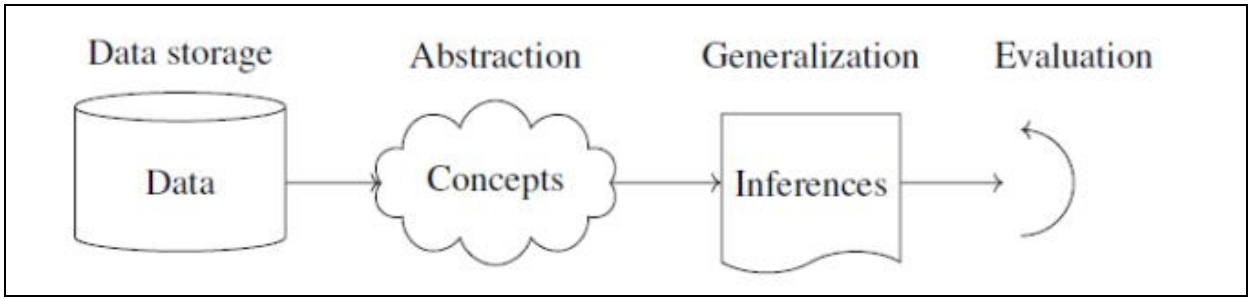


Figure 08. Components Learning Process

2.4 Classes of Machine Learning

Machine Learning algorithms are primarily classified into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning,and Reinforcement learning (Mohammed et al., 2016), as depicted in Figure (09).

The following section provides a brief overview of each learning technique, emphasizing their relevance and applicability in solving real-world problems.

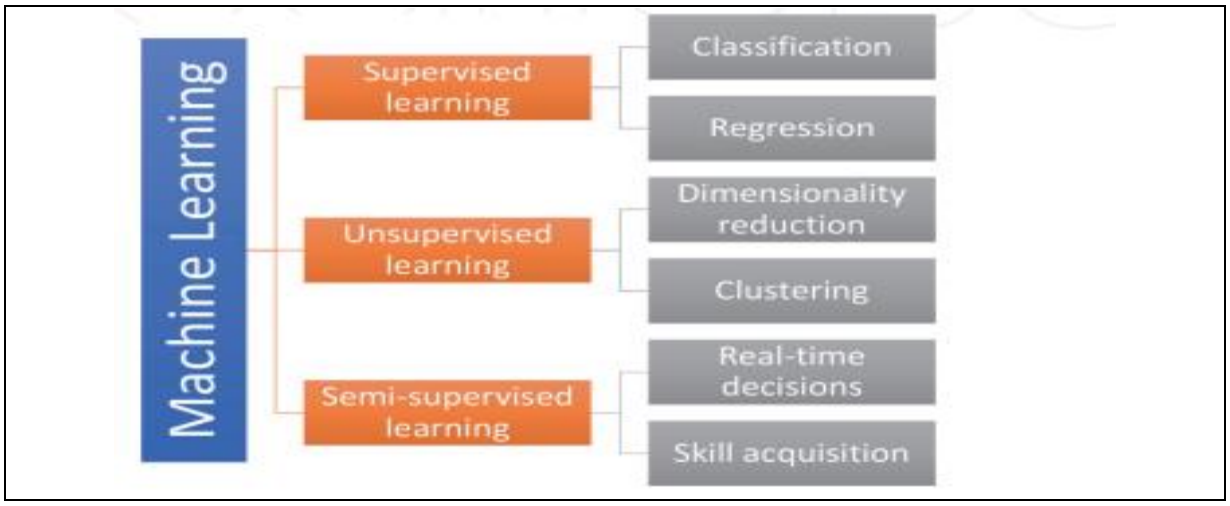


Figure 09. Traditional machine learning types

2.4.1 Supervised learning

Supervised learning is a machine learning approach that involves learning a function that maps inputs to outputs based on sample input-output pairs (Han et al., 2011). It relies on labeled training data and a set of training examples to derive the intended function. This learning method is applied when specific objectives need to be achieved from a defined set of inputs, making it a task-driven approach (Sarker et al., 2020).

The primary tasks in supervised learning include classification, which organizes data into different categories, and regression, which predicts continuous values based on patterns in the data. A practical example of supervised learning is text classification, where a model predicts the class label or sentiment of a text, such as a tweet or a product review.

2.4.1.1 Classification Analysis

Classification is a supervised learning method in machine learning that involves predictive modeling, where a class label is assigned to a given example (Han et al., 2011). Mathematically, it maps a function fff from input variables XXX to output variables YYY as targets, labels, or categories. Classification can be performed on both structured and unstructured data. A common example is spam detection, where emails are classified as either "spam" or "not spam.

- **Binary Classification**

Binary classification involves tasks with two class labels, such as “true and false” or “yes and no” (Han et al., 2011). In such cases, one class may represent the normal state, while the other represents an abnormal state. For example, in a medical test, “cancer not detected” would be the normal state, while “cancer detected” would be considered abnormal. Similarly, in spam detection, “spam” and “not spam” are treated as binary classification labels.

- **Multiclass Classification**

Multiclass classification extends beyond binary classification by involving more than two class labels (Han et al., 2011). Unlike binary classification, there is no strict distinction between normal and abnormal states. Instead, each instance belongs to one of multiple predefined categories. An example is the classification of different types of network attacks in the NSL-KDD dataset, where attacks are categorized into four labels: DoS (Denial of Service Attack), U2R (User to Root Attack), R2L (Root to Local Attack), and Probing Attack (Tavallaee et al., 2009).

- **Multi-label Classification**

Multi-label classification allows each instance to be associated with multiple class labels, making it a generalization of multiclass classification. In these problems, class labels are often hierarchically structured, meaning an instance may belong to multiple categories simultaneously. For example, a news article on Google News may fall under multiple labels such as “city name,” “technology,” or “latest news.” Unlike traditional classification tasks, where class labels are mutually exclusive, multi-label classification allows predicting multiple, non-exclusive categories using advanced machine learning algorithms (Pedregosa et al., 2011).

  - **Core Classification Algorithms**

Many classification algorithms have been developed and are widely applied in various fields. Below, we highlight the core methods that are commonly used across different domains.

- **Decision Tree** (DT): is a widely used non-parametric supervised learning algorithm applied in both classification and regression. It partitions data by recursively applying decision rules, where each internal node represents a feature-based condition, branches indicate possible outcomes, and leaf nodes represent final decisions (Quinlan, 1986). Prominent decision tree algorithms include ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and CART (Breiman et al., 1984).

For splitting nodes, the most commonly used criteria are Gini impurity and entropy for information gain. These can be mathematically expressed as follows:

$$H(x) = -\sum_{i=1}^{n} p(xi)\log^2 p(xi) \quad \text{(Entropy)} \qquad (1)$$

$$\text{Geni E} = 1 - \sum_{i=1}^{c} Pi^2 \quad \text{(Gini Impurity)} \qquad (2)$$

These measures help determine the best attribute to split on at each node, ensuring optimal information gain in classification tasks.

However, while decision trees are powerful, they are prone to overfitting, especially when trained on small or noisy datasets. To mitigate this issue and improve predictive performance, an ensemble learning method called Random Forest (RF) is used.

- **Random Forest:** Random Forest is an ensemble-based approach that constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness. Instead of relying on a single tree, RF employs bootstrap aggregation (bagging) and random feature selection to introduce diversity among the trees, reducing variance and enhancing generalization.

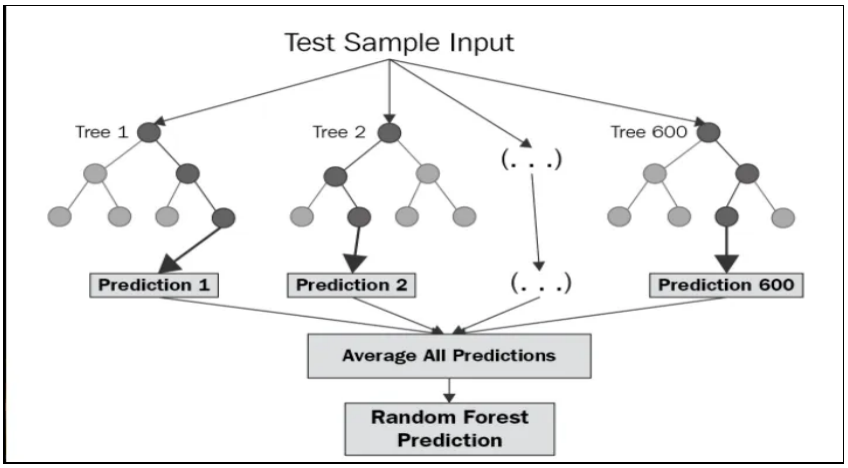


Figure 10. Structure of a Random Forest

- **Extreme Gradient Boosting (XGBoost)**: is an advanced ensemble learning algorithm that constructs a final predictive model by combining multiple decision trees. Similar to

Random Forests, it improves accuracy through ensemble learning, but it further enhances performance by computing second-order gradients of the loss function for more precise optimization. Additionally, XGBoost incorporates L1 and L2 regularization to reduce overfitting and improve generalization. Its ability to efficiently process large datasets makes it a widely used machine learning method (Pedregosa et al., 2011).

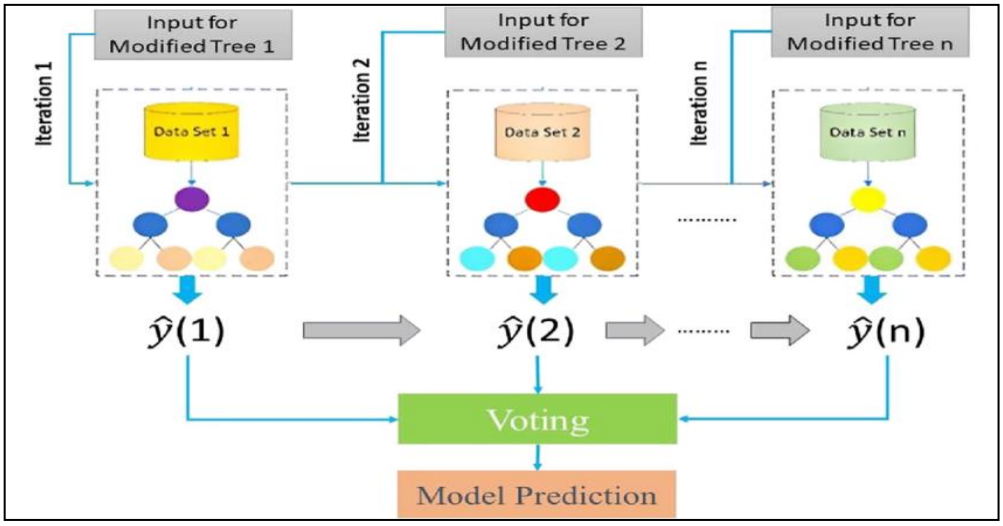


Figure 11. Extreme Gradient Boosting Machine (XGBoost)

Given their ability to handle complex datasets and deliver reliable predictions, I selected Random Forest and XGBoost after evaluating numerous machine learning methods. Both algorithms have been widely applied in various studies and consistently demonstrated high accuracy. For instance, the study "An Implementation of XGBoost and Random Forest Algorithm to Estimate Effective Porosity and Permeability on Well Log Data at Fajar Field, South Sumatera Basin, Indonesia" utilized these models to predict effective porosity and permeability from well log data. The results showed impressive performance, with an R^2 score of 0.90 and a low RMSE of 0.01, highlighting their strong predictive capabilities.

- **The Support Vector Machine (SVM)**: is a powerful machine learning technique primarily used for classification in both supervised and unsupervised settings, with extensions to regression as well (Cortes & Vapnik, 1995). The core concept involves separating data points belonging to different categories by identifying an optimal decision boundary, often referred to as a hyperplane. As illustrated in Figure (12) .this hyperplane divides the feature space such that data points on one side belong to one category, while those on the other side belong to another.

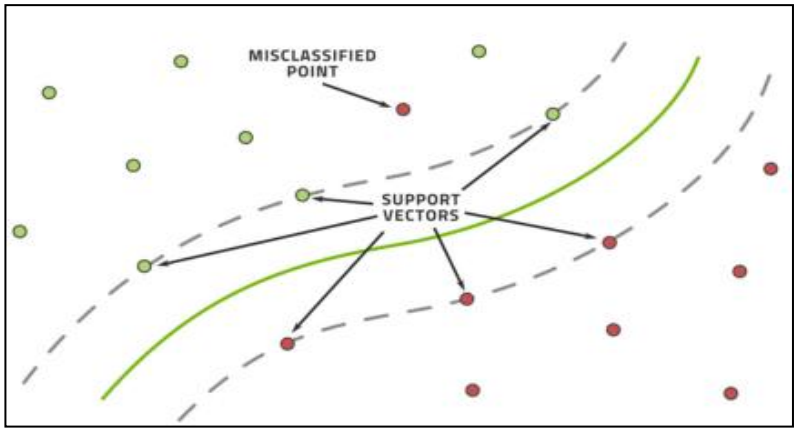


Figure 12. SVM Decision Boundary

SVM determines this boundary based on a subset of critical data points known as support vectors, which define the margin of separation. The objective is to maximize the distance between the hyperplane and the nearest data points from each class, ensuring better generalization. In complex classification tasks, multiple hyperplanes may be required to achieve accurate separation. Additionally, SVM can help identify outliers, as isolated data points in specific sections of the feature space may indicate anomalies.

Recognizing its strong classification capabilities and efficiency in handling complex datasets, I selected Support Vector Machine (SVM) as a key algorithm. Its effectiveness has been well demonstrated in research, such as the study "Lithology Identification by Support Vector Machine Using Well Logging Data," where SVM achieved an impressive 85.1% prediction accuracy, outperforming back-propagation neural networks. This further highlights its reliability in subsurface modeling and geoscience applications.

2.4.1.2 Regression Analysis

Encompasses a range of machine learning techniques designed to predict a continuous outcome variable (y) based on one or more predictor variables (x). The key distinction between regression and classification lies in their objectives—while classification assigns data points to distinct categories, regression estimates a continuous value. Figure (13) illustrates this fundamental difference. However, some machine learning algorithms can be adapted for both tasks, leading to occasional overlaps between classification and regression models.

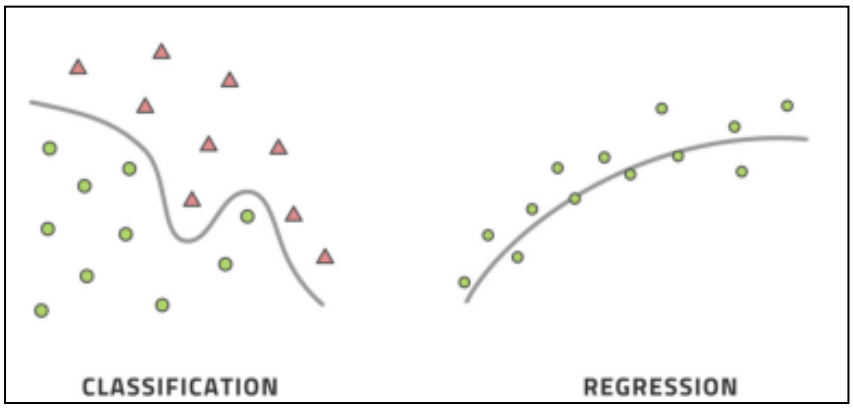


Figure 13. The difference between classification and regression methods

- **Core Regression Algorithms**

Some of the familiar types of regression algorithms are linear, polynomial, lasso and ridge regression, etc., which are explained briefy in the following

- **Simple and multiple linear regression**: Linear regression is one of the most widely used machine learning techniques for supervised learning, designed to predict a continuous dependent variable (Y) based on one or more independent variables (X), which can be either continuous or discrete. This method establishes whether a linear relationship exists between the dependent and independent variables by fitting the best-fit straight line, known as the regression line (Han, J., Pei, J., & Kamber, M., 2011).

Linear regression is categorized into simple linear regression, where only one independent variable is used, and multiple linear regression, which extends the model to include two or more predictor variables. It is mathematically defined as:

$$y = a + b\,x + e \qquad (3)$$

where (a) is the intercept, (b) is the slope of the line, and (e) is the error term. This equation enables the prediction of the target variable based on the given predictor (s). In the case of multiple linear regression, the model generalizes to:

$$y = a + b_1\,x_1 + b_2\,x_2 + \cdots + b_n\,x_n + e \qquad (4)$$

where multiple independent variables contribute to predicting Y. Due to its simplicity, interpretability, and effectiveness, linear regression remains a fundamental technique in various fields, from finance to geoscience.

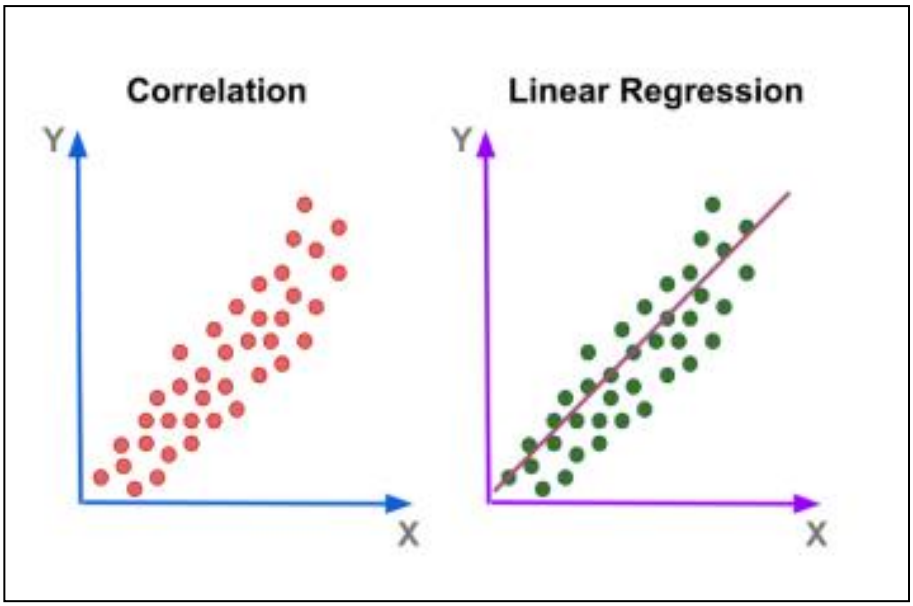


Figure 14. Linear regression and correlation analysis

- **Polynomial regression**: a type of regression analysis where the relationship between the independent variable xxx and the dependent variable yyy follows a polynomial function rather than a linear one. It extends linear regression by incorporating polynomial terms of degree nnn in xxx. The mathematical formulation of polynomial regression is derived from the linear regression equation (which corresponds to a polynomial of degree 1) and is expressed as follows:

$$Y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n \qquad (5)$$

where (y) represents the predicted output, (x) is the independent variable, $b_0, b_1, b_2, \ldots\ldots, b_n$ are the regression coefficients. In essence, polynomial regression is employed when the data does not exhibit a linear trend but instead follows a higher-degree polynomial pattern, allowing for a more accurate representation of complex relationships.

The comparison between MLR and PR demonstrated the superior performance of PR in capturing nonlinear relationships. As shown in "Application of Multiple Linear and Polynomial Regression in the Sustainable Biodegradation Process of Crude Oil" by Ajona et al. (2023), PR achieved a higher R^2 value of 0.863, significantly outperforming MLR, which had an R^2 of 0.495. This indicates PR's greater ability to model complex dependencies, reinforcing its relevance for this application.

LASSO and Ridge regression are powerful techniques for handling high-dimensional datasets by reducing overfitting and improving model interpretability.

- LASSO (Least Absolute Shrinkage and Selection Operator) uses L1 regularization, which penalizes the absolute values of regression coefficients, effectively shrinking some to zero. This makes it useful for feature selection by eliminating less relevant predictors.

Ridge regression, on the other hand, applies L2 regularization, penalizing the squared magnitude of coefficients. Unlike LASSO, it does not set coefficients to zero but reduces their impact, making it effective for handling multicollinearity.

In summary, LASSO is ideal for feature selection, while Ridge is better suited for correlated predictors, ensuring more stable and reliable models (Pedregosa et al., 2011).

Table 01. Comparison of Regression Techniques

Regression Type	Key Takeaways
Least-Squares Linear Regression	Simple and effective, but may lack accuracy.
Ridge Regression	Reduces model complexity by penalizing less influential features.
Lasso Regression	Identifies key features by shrinking less important ones to zero.
Polynomial Regression	Solves non-linear regression problems while maintaining a linear framework.

2.4.2 UnSupervised learning

Unlike supervised learning, which uses labeled data, unsupervised learning analyzes unlabeled datasets to identify patterns and relationships. It requires large datasets to detect hidden structures without predefined categories (Bantan et al., 2020).

As shown in Figure (15), key challenges in unsupervised learning include clustering, association, anomaly detection, and autoencoders, which help categorize and interpret complex data.

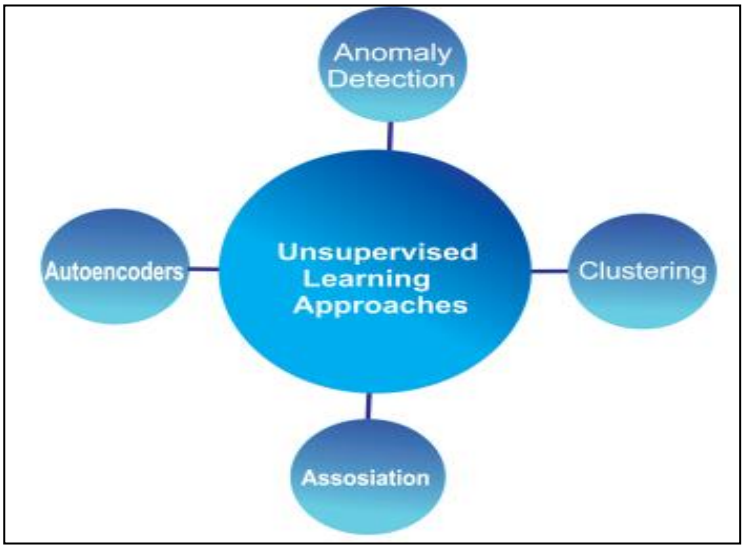


Figure 15 . Types of Unsupervised Learning

2.4.2.1 Types of UnSupervised learning

- **Clustering:** Clustering, or cluster analysis, is the process of grouping data points based on similarity. It includes partitioning, hierarchical, overlapping, and probabilistic clustering (Kabir & Luo, 2020) as shown in Figure (16)**.**

Clustering methods vary:

- Exclusive Clustering**:** Each data point belongs to only one cluster, such as in K-Means**.**
- Hierarchical Clustering**:** Forms a hierarchy of clusters using agglomerative (bottom-up) or divisive (top-down) approaches (Ameur-Zaimeche et al., 2020).
- Overlapping Clustering**:** Allows data points to belong to multiple clusters with varying degrees of membership (Saba & Ngepah, 2022).

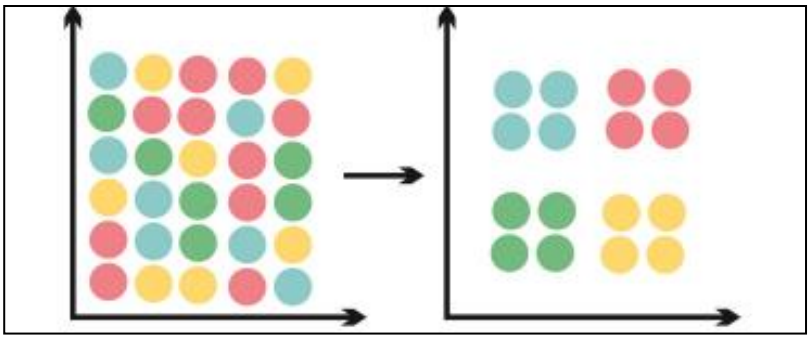


Figure 16. Example of Clustering

- **Association:** Association Rule Learning (ARL) identifies relationships between variables in large datasets. It is commonly used in recommendation systems, shopping cart analysis, and web usage mining. ARL relies on support (how often a pattern occurs) and confidence (how frequently an if/then relationship holds).

- **Anomaly Detection:** Anomaly detection finds outliers in data that deviate from normal patterns. It is used in fraud detection, cybersecurity, and sensor monitoring. Atypical network traffic or unexpected behavior in data can indicate security threats or system failures.
- **Autoencoders:** Autoencoders are neural networks that learn efficient data representations through compression and reconstruction. They identify patterns by reducing data dimensions, making them useful for anomaly detection and feature extraction. Figure (17) illustrates the autoencoder process.

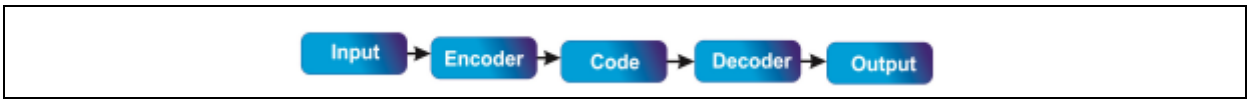


Figure (17). Autoencoder Components (Recherchegate)

2.4.2.2 Unsupervised Learning Algorithms

Several algorithms are commonly used in unsupervised learning, each serving different purposes:

- **K-means clustering** : a fast, robust, and simple algorithm that performs well when data sets are well-separated (MacQueen et al., 1967). It assigns data points to clusters by minimizing the squared distance between points and their respective centroids. The algorithm identifies k centroids and assigns each data point to the nearest one, aiming to keep centroids compact. However, due to its random initialization, results can vary. Additionally, K-means is sensitive to outliers, as extreme values can heavily influence the mean. A more robust variant, K-medoids clustering, improves resistance to noise and outliers. (Ameur Zaimeche et al. 2014)

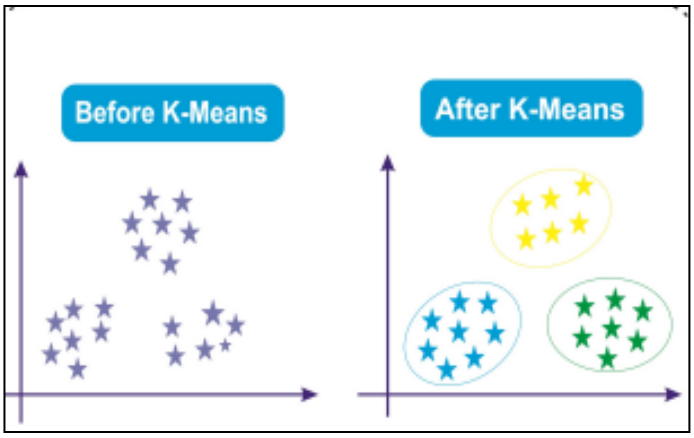


Figure (18). Workflow of K-mean Clustering

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**:

a widely used density-based clustering algorithm in data mining and machine learning. Unlike parametric methods, DBSCAN separates high-density regions from low-density ones without requiring a predefined number of clusters. A point is assigned to a cluster if it is surrounded by many other points from the same cluster. This method effectively identifies clusters of various shapes and sizes, even in noisy datasets with outliers. While K-means is computationally faster, DBSCAN excels at detecting high-density regions and is more robust to outliers.

- **Gaussian Mixture Models (GMMs)**: offer a probabilistic approach to clustering by assuming that data points are generated from a mixture of Gaussian distributions. Unlike K-means, which assigns points to the nearest cluster, GMMs account for uncertainty by providing the probability of a data point belonging to each cluster. The Expectation-Maximization (EM) algorithm is typically used to estimate the parameters of these Gaussian distributions. GMMs are particularly useful for handling complex, non-linear data distributions and offer greater flexibility compared to K-means clustering.

2.4.3 Semi-Supervised learning

A hybrid approach that combines elements of both supervised and unsupervised learning, utilizing both labeled and unlabeled data. It bridges the gap between fully supervised and unsupervised methods, making it particularly useful in scenarios where labeled data is scarce while large amounts of unlabeled data are available. The primary objective of semi-supervised learning is to enhance prediction accuracy beyond what is achievable using only labeled data. Common applications include machine translation, fraud detection, data labeling, and text classification.

Some of the popular semi-supervised learning algorithms include:

- **Self-training**: Uses a supervised model trained on labeled data to predict labels for the unlabeled data, iteratively expanding the training set.
- **Co-training**: Utilizes two classifiers trained on different feature subsets to label unlabeled data for each other.
- **Graph-based methods**: Leverage graph structures to propagate labels from labeled to unlabeled data based on data similarity.
- **Transductive Support Vector Machines (TSVMs)**: Extend traditional SVMs to utilize both labeled and unlabeled data, optimizing decision boundaries.
- **Generative models**: Employ probabilistic models like Gaussian Mixture Models (GMMs) to estimate data distributions and infer labels.

2.4.4 Reinforcement Learning (RL)

Enables an agent to learn optimal actions through trial and error, guided by rewards. It is modeled as a Markov Decision Process with key components: agent, environment, rewards, and policy.

RL is divided into Model-Based (uses an environment model) and Model-Free (learns directly from interactions). Examples of Model-Free RL include Q-Learning, Deep Q Networks (DQN), Monte Carlo Control, and SARSA. The key difference is that Model-Based RL relies on a policy network, while Model-Free RL does not.

Next, we explore popular RL algorithms.

Monte Carlo : s refer to a broad category of computational algorithms that use repeated random sampling to obtain numerical results. As highlighted by Kaelbling, Littman, and Moore (1996), these methods leverage randomness to solve problems that are inherently deterministic. They are particularly useful in optimization, numerical integration, and probability distribution modeling.
Q-learning: is a model-free reinforcement learning algorithm that determines the best actions by maximizing expected rewards. It adapts to stochastic environments without needing a predefined model.
Deep Q-learning: enhances Q-learning by using a neural network to estimate Q-values for possible actions. It is effective in complex environments where traditional Q-learning struggles with large state-action spaces.

3. Deep learning:

Deep Learning (DL) is a subset of Machine Learning (ML) that uses artificial neural networks with multiple layers (deep neural networks) to automatically learn patterns and representations from large amounts of data. It is inspired by the structure and function of the human brain and excels in tasks such as image and speech recognition, natural language processing, and autonomous decision-making. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can learn hierarchical features and complex patterns, making them highly effective for solving real-world problems.

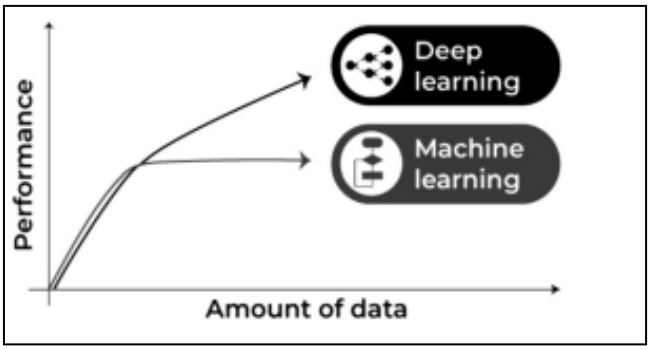


Figure19. Machine learning and deep learning performance in general with the amount of data

3.1 Artificial Neural Networks (ANNs): Neural networks in deep learning consist of multiple interconnected layers that process and transform data, drawing inspiration

from the human brain. They learn from large volumes of data and improve through training using complex algorithms. The concept of artificial neural networks (ANNs) dates back to the 1940s with the development of threshold logic, a mathematical model that laid the foundation for neural network research. However, due to limited computing power, ANNs lost popularity for several decades. With advancements in technology and more efficient algorithms, neural networks regained interest and became widely implemented, leading to breakthroughs in deep learning applications such as image recognition, speech processing, and decision-making.

3.1.1 Structure of Neural Networks

The structure of an artificial neural network is similar to that of a biological neuron. Biological neurons consists of Dendrites which acts as input and is responsible for the collection of external stimuli. This Information (stimuli) passes through a thin long strand, Axon, which acts as the output. There is one other thing called the Synapse which converts the information from Axon into an electrical signal.

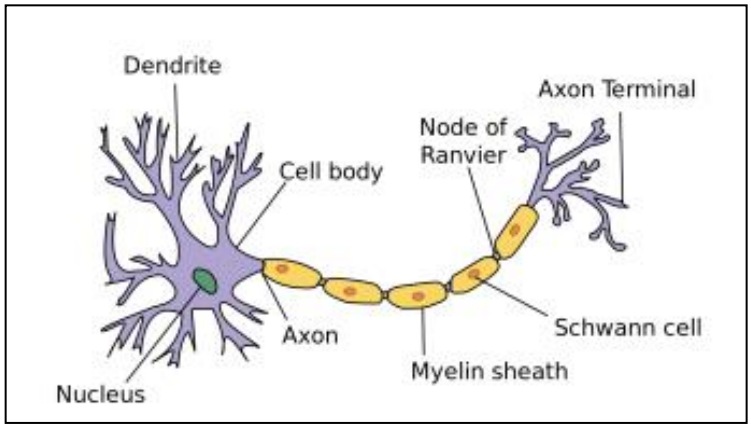


Figure 20. Biological Neural Network present in Human Body

An artificial neural network (ANN) consists of numerous interconnected units (neurons) that work together to solve specific problems. The connections between neurons are assigned **weights**, which determine their influence on each other, while the **activation value** represents the output of each neuron. A neural network is typically composed of three main layers Figure (21)

- **Input Layer:** This layer receives external information or data, which is then passed to the next layers for processing. The number of neurons in the input layer corresponds to the dimensions of the input data.
- **Hidden Layers:** These intermediate layers perform complex transformations on the data by applying weighted connections between neurons. A network can have multiple hidden layers, and increasing their number enhances the network's ability to learn intricate patterns.
- **Output Layer:** This layer provides the final prediction or classification result. The number of neurons in the output layer corresponds to the number of target classes in a classification task.

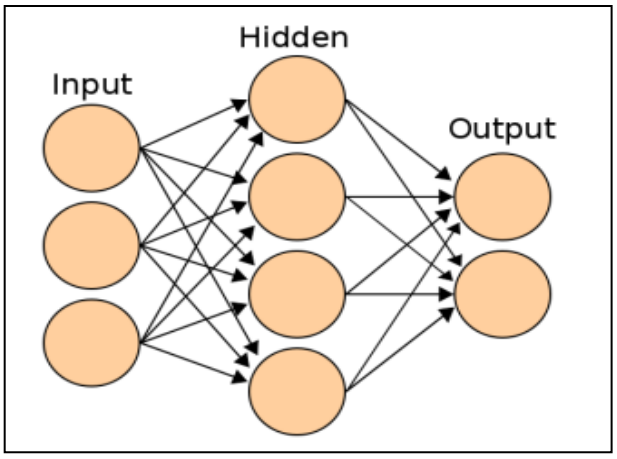


Figure 21. A three layered Artificial Neural Network

Each layer consists of multiple neurons connected through weighted links, allowing the network to progressively extract and learn features from the data. In general, deeper networks with more hidden layers improve learning performance by capturing complex relationships within the input data.

3.2 ALGORITHMS

Convolutional Neural Network (CNN): s a type of feed-forward neural network designed primarily for image recognition and processing. Unlike traditional neural networks, CNNs are structured to efficiently analyze visual data with minimal pre-processing. They are a subclass of deep artificial neural networks that utilize convolution operations to extract meaningful features from images.

CNNs follow a unidirectional flow of information, meaning there are no loops or cycles within the network. The key component of CNNs is the convolution operation, which is mathematically similar to cross-correlation but differs as it flips across its height. Due to the associative property of convolution, it plays a crucial role in extracting spatial hierarchies of features from input images.

A CNN typically consists of three main types of layers:

- **Convolutional Layer :** Applies filters to the input data to detect patterns such as edges, textures, and shapes.
- **Pooling Layer :** Reduces the dimensionality of feature maps, retaining essential information while making the network computationally efficient.
- **Fully Connected Layer:** Combines extracted features and makes the final classification or prediction.

By stacking multiple convolutional and pooling layers, CNNs can learn complex representations of images, making them highly effective in computer vision tasks such as object detection, facial recognition, and medical image analysis.

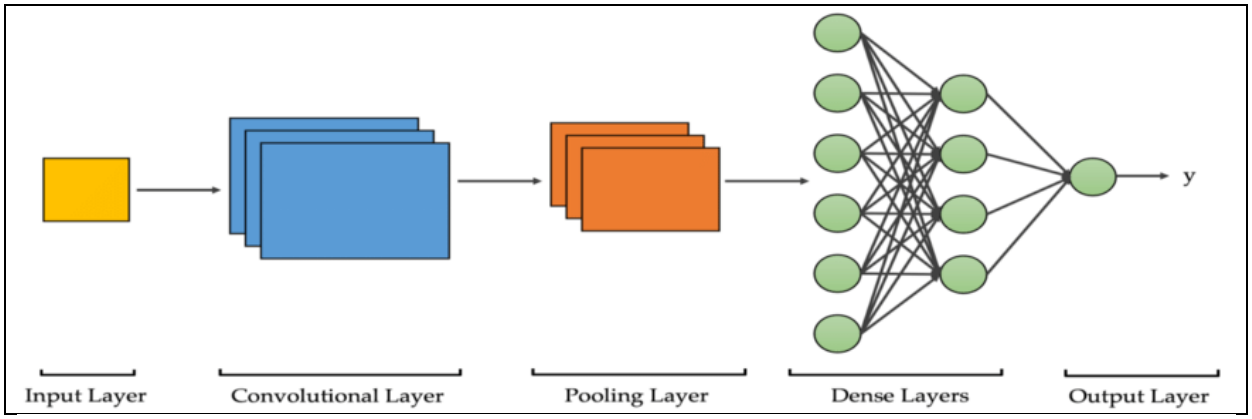


Figure 22. Basic convolutional neural network structure ((LeCun, 1989)

A Recurrent Neural Network (RNN): is a type of neural network that forms directed cycles or loops, allowing it to process sequential data by retaining information from previous computations. This memory mechanism enables RNNs to recognize patterns over time, making them suitable for tasks involving temporal dependencies. Unlike feedforward networks, RNNs process inputs dynamically, where each output depends on both the current and past inputs. The network's structure allows information to flow across different time steps, making it effective for applications such as speech recognition, language processing, and time-series analysis.

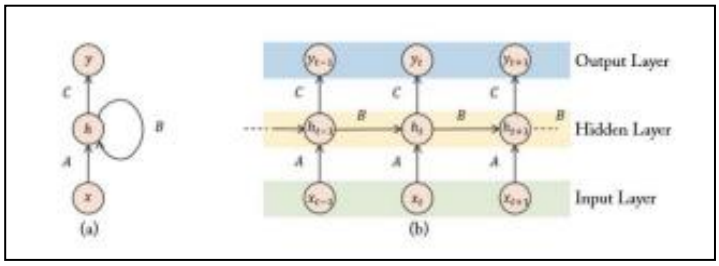


Figure 23. (a) A simple RNN structure with a feedback loop (b) An unfolded RNN at different time-steps

4. Advances and Future Directions in Machine Learning (PIML)

Physics-Informed Machine Learning (PIML) is an advanced approach that integrates machine learning models with fundamental physical principles to enhance predictive accuracy and generalization. Unlike conventional data-driven models that rely purely on empirical data, PIML incorporates prior knowledge from physical laws, such as conservation laws and differential equations, ensuring model predictions remain consistent with established scientific theories (Hao et al., 2021).

The key advantage of PIML is its ability to improve model interpretability and efficiency (Khassaf et al., 2025), particularly in domains where data is scarce or expensive to obtain, such as fluid dynamics, materials science, and climate modeling. By embedding physics-based constraints into the learning process, PIML reduces the dependency on large datasets while enhancing model robustness.

Several methodologies are commonly used in PIML, including:

- **Physics-Informed Neural Networks (PINNs):** These networks solve partial differential equations (PDEs) by incorporating physics-based loss functions, allowing them to generalize well even with limited data
- **Hybrid Models:** A combination of traditional machine learning models and physics-based simulations, improving both accuracy and reliability.
- **Regularization Techniques:** Physics constraints are added as penalty terms in the loss function, guiding the training process to ensure compliance with physical laws.

Future advancements in PIML aim to enhance computational efficiency and extend applications to more complex, real-world problems. The introduction of novel optimization techniques, such as bi-level physics-informed neural networks for PDE-constrained optimization using Broyden's hypergradients, has further improved PIML's ability to solve

Intricate scientific challenges.As machine learning continues to evolve, PIML is expected to play a pivotal role in scientific discovery and industrial innovation.

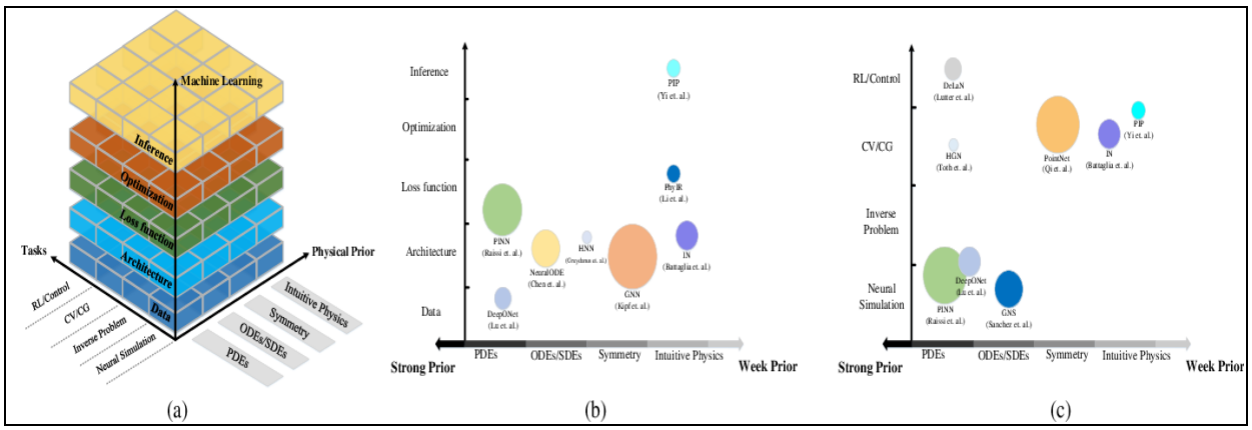


Figure 24. Overview of physics-informed machine learning (Hao, Liu, Zhang, Ying, Feng, Su, & Zhu, 2023).

Chapter 03

Review of Real-Time Petrophysical Prediction

Introduction

The precise estimation of petrophysical parameters—namely porosity, permeability, and fluid saturation—constitutes a fundamental aspect of reservoir characterization and plays a pivotal role in maximizing hydrocarbon recovery. These parameters serve as the cornerstone for critical operational decisions in well placement, drilling optimization, reservoir simulation, and production forecasting. Traditionally, such estimations have relied on the interpretation of well logs and core sample analyses. However, these conventional techniques are increasingly hindered by several limitations, including high operational costs, environmental sensitivities, tool calibration errors, and delayed data availability, particularly in complex geological formations.

In recent years, the petroleum industry has witnessed a paradigm shift towards data-driven, machine learning (ML) methodologies. ML models, known for their ability to model non-linear and high-dimensional relationships, offer a promising alternative to traditional deterministic approaches. While well logs have historically been the primary input for ML-based petrophysical modeling, their static nature limits real-time applicability. As exploration and production environments evolve toward real-time decision-making and automation (Kermia et al., 2025) the integration of dynamic data sources such as real-time drilling data (DD) and gas-while-drilling (GWD) has emerged as a transformative advancement.

Previous reviews have laid the groundwork for understanding ML applications in this domain. Smith et al. (2019) explored ML-based permeability estimation using conventional logs, while Johnson et al. (2020) focused on porosity prediction. More comprehensively, Zhang et al. (2023) categorized ML applications into five core domains: porosity prediction, permeability estimation, fluid saturation modeling, lithology classification, and well log interpretation. However, a critical gap persists in the literature: these studies largely overlook the operational limitations of static log-based models and fail to address the growing trend toward hybrid, real-time AI frameworks.

Operational reports further underscore the shortcomings of relying solely on well logs. According to Halliburton (2020) and Schlumberger (2019), well log interpretations are frequently compromised by geological heterogeneity, borehole condition variability, and tool malfunctions. These factors introduce significant uncertainty and reduce the reliability of petrophysical estimates. In contrast, real-time DD and GWD data provide continuous, high-frequency measurements such as rate of penetration (ROP), weight-on-bit (WOB), torque, and subsurface gas composition. These dynamic inputs enable the development of adaptive ML models capable of generating real-time predictions and facilitating in-the-moment operational adjustments.

Recent empirical investigations validate this transition. For instance, Ameur-Zaimeche et al. (2022) developed an inverse intelligent ML model using GWD data calibrated against core porosity, where extreme learning machines (ELM) outperformed support vector regression and random forest regression, achieving a correlation coefficient of 0.929. Similarly, Sharma et al. (2023) demonstrated that hybrid models integrating DD, GWD, and conventional logs significantly improved permeability prediction accuracy, further emphasizing the superiority of real-time learning systems.

This review seeks to provide a comprehensive synthesis of the evolving landscape of ML in petrophysical parameter prediction, with the following specific objectives: (1)) to conduct a

temporal keyword analysis using VOSviewer to examine trends in machine learning applications, research collaborations, and citation patterns (2) to perform a publication trends analysis to trace the progression of real-time ML applications in petrophysical prediction, highlighting key shifts in research focus, data integration, and industry adoption over time (3) To catalog and evaluate common ML algorithms (4) Focus on outputs, gaps, and multi-target prediction

By integrating insights from both historical and cutting-edge methodologies, this study aims to underscore the transformative potential of real-time ML frameworks in reshaping petrophysical analysis. Ultimately, the research advocates for a future where AI-driven, adaptive systems enhance reservoir understanding, reduce operational risks, and enable more efficient, intelligent hydrocarbon exploration and production.

1. Literature Review – Petrophysical parameters

Petrophysical parameters such as porosity, permeability, fluid saturation, and bulk density are fundamental to the assessment and management of subsurface reservoirs. These properties directly impact reservoir behavior and ultimately determine the efficiency of hydrocarbon extraction.

Porosity, defined as the ratio of void space to total rock volume, serves as a critical indicator of a reservoir's storage capacity. It influences the amount of recoverable hydrocarbons and provides a baseline for evaluating reservoir potential (Tiab & Donaldson, 2015). Permeability, by contrast, quantifies the ability of rock to transmit fluids, making it essential for predicting flow capacity and informing well deliverability forecasts (Bjørlykke, 2010). Fluid saturation describes the distribution of fluids within the rock matrix and plays a key role in estimating recoverable volumes, influencing both production rates and reservoir management strategies (Dandekar, 2006). Bulk density, closely linked to porosity and mineral composition, offers insight into rock structure and mass, contributing to porosity estimation and overall reservoir quality assessment (Ellis & Singer, 2007).

When considered together, these parameters provide a comprehensive characterization of reservoir properties under varying operational conditions. Accurate prediction of these variables is essential not only during initial exploration but also for continuous reservoir management and optimization throughout the field's lifecycle (Lucia, 2003). Their role is especially prominent in strategic decisions such as well design, production planning, and enhanced oil recovery initiatives.

To track the evolution of research focused on predicting these parameters through ML, a temporal keyword analysis was conducted using VOSviewer. The analysis, based on literature retrieved from Google Scholar and Scopus spanning 2019 to 2025, produced a network map (Figure 25), that visually represents the relationships and prominence of key research themes over time.

The keyword map highlights several significant trends. A persistent and growing area of interest lies at the intersection of deep learning algorithms with core principles of rock physics and reservoir estimation—demonstrating an industry-wide move toward more sophisticated modeling techniques. Permeability prediction stands out as a consistently explored domain, closely associated with porosity due to their interdependence in fluid flow modeling. Notably,

terms such as "drilling," "realtime," and "optimization" appear with increasing frequency and warmer color tones in recent years, signaling a rising emphasis on real-time applications.

Beyond these central themes, the analysis also identifies a broader diversification of ML usage across other petrophysical variables, including water saturation, density, and lithology. This indicates an expanding scope in how ML is being employed to model a wide array of subsurface characteristics. The visualization not only underscores the centrality of ML in modern reservoir studies but also reveals the interconnectedness of the field's evolving focus areas, from static predictions to dynamic, real-time systems.

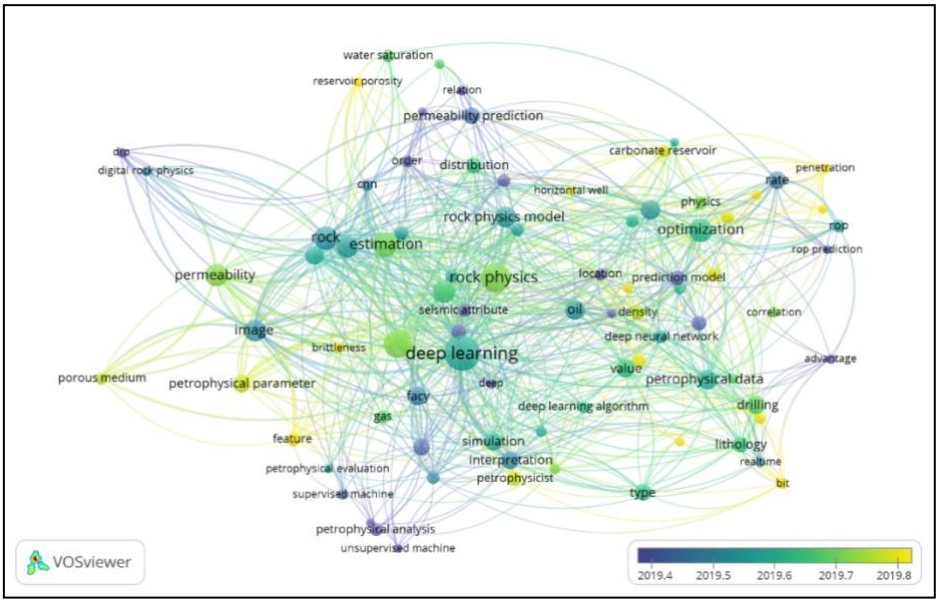


Figure25.Temporal Keyword Analysis of Machine Learning Applications in Petrophysical Parameter Prediction (2019–2025)

2. Publication Trends Analysis : From Exploration to Real-Time Transformation

Building on the temporal keyword analysis, which highlighted the growing emphasis on real-time data integration and intelligent prediction systems in petrophysical workflows, this section investigates how such themes have evolved in the academic literature. Research on real-time petrophysics prediction has experienced continuous development in recent years, as evidenced by the database table (02) documenting the number of studies published annually. A curated dataset of 37 studies published between 2019 and 2025 (Figure 26) reveals a nonlinear yet insightful research trajectory. This trend reflects not only the development of machine learning (ML) methodologies but also changes in data accessibility, computational infrastructure, and industry openness to digital solutions.

➢ Early Stages (2019–2020)

In 2019 and 2020, real-time ML applications in petrophysical prediction were limited. Research primarily relied on conventional data sources—wireline logs and core data—due to limited access to live-streamed drilling data. During this phase:

- There was fragmented and unstructured mud logging and gas data.

- Industry resistance to complex ML models persisted, primarily due to concerns over transparency and explainability. As a result, real-time implementations remained largely theoretical, with few publications emerging.

➢ **Data Interpretability and Adoption (2021)**

By 2021, advancements in computational power, data acquisition, and model interpretability spurred greater academic engagement. Techniques like SHAP and LIME improved model transparency, reducing industry resistance. Experimental workflows began integrating drilling and gas data for real-time simulations.

Notably, Tian et al. (2021) introduced a hybrid ML approach for permeability prediction, highlighting the growing sophistication of reservoir characterization models. This period also saw early efforts to incorporate physics-informed constraints into data-driven models, enhancing alignment with geological reasoning and operational reliability.

➢ **Peak Growth (2022) in Real-Time ML Applications**

2022, the number of publications peaked, driven by the continued evolution of machine learning in reservoir characterization. The surge in studies was fueled by stronger collaborations between academia and service companies, which provided access to previously restricted field data, as well as the growing acceptance of data-driven decision-making in high-risk operations. These factors significantly contributed to the rise in studies focusing on real-time prediction of petrophysical parameters using active drilling data.

Notable contributions from this period include Anifowose et al. (2022), who explored the use of advanced mud gas data to predict reservoir porosity, and Gamal et al. (2022), who introduced an interpretable gradient boosting framework for permeability estimation, demonstrating strong performance across a range of stratigraphic formations.

➢ **Refinement Phase of Real-Time Modeling (2023-2024)**

In 2023 and 2024, research increasingly focused on addressing specific geological challenges, such as fractured carbonates and tight sands, and integrating multi-source data like mud gas, drilling mechanics, and annular pressure trends. Key efforts aimed to improve model interpretability by incorporating physics-based constraints and uncertainty quantification. Notably, Alhowaish et al. (2023) highlighted operational challenges and proposed solutions for real-time porosity prediction in complex reservoirs. Although publication volume slightly decreased, research quality and specialization improved, focusing on more geologically nuanced environments and tailored machine learning frameworks (Ameur-Zaimeche, Kechiched, & Aouame, 2021) . This shift reflects the maturation of real-time modeling, with a growing emphasis on regional geological contexts

➢ **Future Trends Outlook**

2025, real-time machine learning applications for petrophysical prediction will continue to advance, particularly through the integration of drilling and gas data. Research will focus on refining frameworks for more accurate, interpretable real-time decision-making, with an emphasis on uncertainty quantification and complex reservoir scenarios. Boutaghane et al. (2025) introduced an interpretable boosting framework for bulk density prediction, pushing the

frontier of operational ML. Although publication volume is low in 2025, this is likely due to the early stage of the year and typical delays in model development, peer review, and publication.

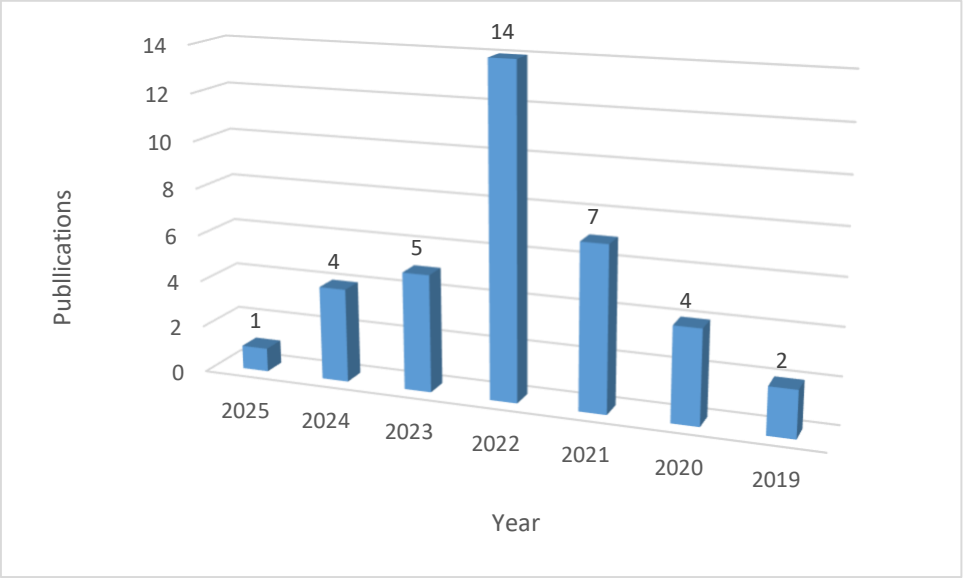


Figure 26. Temporal distribution of relevant research articles from (2019 / 2025)

3 Algorithmic Evolution and Usage Trends in Real-Time Petrophysical Modeling

To contextualize the shift toward real-time, data-driven frameworks, it is essential to trace how machine learning (ML) applications and algorithm choices have evolved over time in petrophysical parameter prediction. The development trajectory reflects three key phases, each characterized by shifts in methodological focus, data accessibility, and operational priorities.

- **Phase One (2019–2021): Exploring Initial Models and Approaches**

This period marked the early exploration into the viability of machine learning for real-time petrophysical parameter prediction. Researchers primarily focused on evaluating the feasibility of Artificial Neural Networks (ANN) and Deep Neural Networks (DNN) as baseline models for data-driven insights into geological and reservoir properties.

Studies such as Zolotukhin et al. and Arigbe et al. laid the groundwork by using neural networks to estimate parameters like permeability. In 2020, algorithmic diversity increased with the introduction of Genetic Algorithms (GA) and sequential models like Multilayer Long Short-Term Memory (MLSTM), as explored by Al-AbdulJabbar et al., Wang et al., and Chen et al.

By 2021, the research ecosystem broadened significantly. Algorithms such as Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVM), and various forms of Neural Networks (NN) began to gain traction. This expansion into tree-based models and margin-based classifiers brought improvements in prediction robustness and interpretability. Studies from Tian et al., Gamal et al., Farouk et al., and others reflected this shift.

The algorithms used in this phase belonged primarily to two pillars: Machine Learning (ML) and Deep Learning (DL**).** DL models like ANN and DNN were instrumental in extracting complex patterns via hierarchical feature learning, while traditional ML models such as RF and SVM provided strong performance with increased explainability. However, DL adoption remained selective, often constrained by computational resources and data availability.

- **Phase Two (2022–2023): Expanding Algorithmic Horizons**

In this phase, the research trajectory shifted toward improving model accuracy and expanding the suite of algorithms. The introduction of Extreme Learning Machines (ELM), Random Forest Regressors (RFR), Multilayer Perceptrons (MLPNNs), and General Regression Neural Networks (GRNN) marked a move toward lighter and faster models. These models offered rapid training and inference times, a key factor for real-time systems.

Simultaneously, Gradient Boosting Machines (GBMs) such as XGBoost (XGB) became increasingly prevalent. Studies by Mohammadian et al. and Qalandari et al. highlighted XGB's efficiency in handling non-linearity and its resilience to overfitting. In parallel, Kamali et al. introduced Support Vector Regression (SVR), Polynomial Regression (PR), and the Group Method of Data Handling (GMDH)—a flexible, self-organizing approach capable of modeling complex relationships.

By 2023, algorithms like K-Nearest Neighbors (KNN), RF, SVR, and XGB were frequently employed, as shown in works by Ouladmansour and Sharma. The growth in DL usage included Convolutional Neural Networks (CNNs) for spatial pattern recognition and Recurrent Neural Networks (RNNs) for modeling time-series signals from drilling data and mud gas.

This phase emphasized hybridization—blending ML and DL techniques to strike a balance between model precision, speed, and interpretability. Researchers could now tailor algorithm choice to specific challenges: KNN for localized behavior, SVR for smoother generalization, and CNN/RNN for sequential and multidimensional data analysis.

- **Phase Three (2024–Present): Integration with Modern Technologies and Accuracy Enhancement**

The current phase focuses on operationalizing AI for field-ready, real-time petrophysical parameter estimation. There's a strong trend toward integrating ML algorithms with modern computing platforms and sensor systems, enhancing both model accuracy and deployment efficiency.

Studies by Hassaan et al. and Anifowose et al. in 2024 demonstrated that mature algorithms like RF, SVM, DT, and Gradient Boosting Regressors (GBRs) remain central due to their balance between performance and interpretability. Ensemble methods, especially boosting-based frameworks, continue to outperform in predictive stability and generalization, particularly in noisy or limited-data environments.

Advanced DL architectures (Figure 27) such as CNNs and RNNs are now being used more confidently for real-time processing. Their ability to learn from high-frequency, multivariate drilling and mud gas signals has proven particularly useful in dynamic formation evaluation.

In this stage, algorithm selection is becoming more strategic—focusing not only on accuracy but also on transparency, adaptability, and responsiveness. Models are now chosen based on formation complexity, real-time data fidelity, and computational constraints. The integration of hybrid and ensemble models points to a more mature, application-driven era in real-time petrophysical analysis.

As early studies in 2025 indicate, the field appears to be moving toward interpretable yet high-performance models—particularly those leveraging gradient boosting and modular neural architectures. These are poised to dominate future workflows, especially as real-time datasets grow richer and decision-making systems demand greater trustworthiness and transparency.

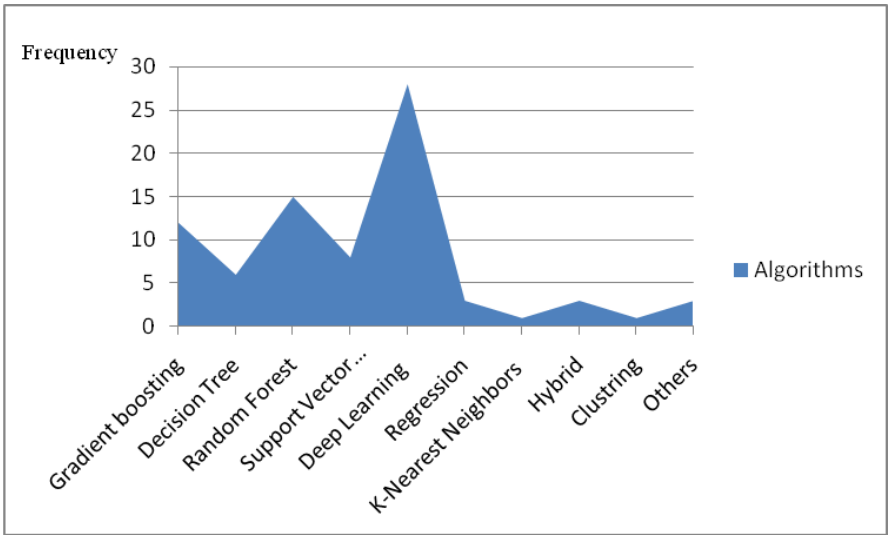


Figure 27. Frequency of algorithms used between (2019 2025)

4 .Distribution of Predicted Petrophysical Outputs in Real-Time Machine Learning

Across the current body of research, clear trends emerge regarding the focus of real-time machine learning models on specific petrophysical outputs. Porosity stands out as the most frequently predicted parameter, reflecting its foundational role in reservoir evaluation and its compatibility with real-time data sources like drilling mechanics and mud gas measurements. It accounts for more than half of the examined studies. Permeability follows, targeted in approximately 30% of cases, due to its direct link to reservoir deliverability and production planning. Bulk density appears less frequently, featured in around 10% of works, typically alongside porosity for integrated rock characterization. Fluid saturation, despite its importance, is addressed in only 2% of studies—likely due to the difficulty in accurately modeling saturation without resistivity-based inputs. This uneven focus highlights a tendency to prioritize parameters with more accessible proxies in real-time workflows, while also pointing to underexplored opportunities in multi-target prediction and saturation modeling. The accompanying pie chart Figure (28) visualizes this distribution and underscores the current research emphasis and remaining gaps.

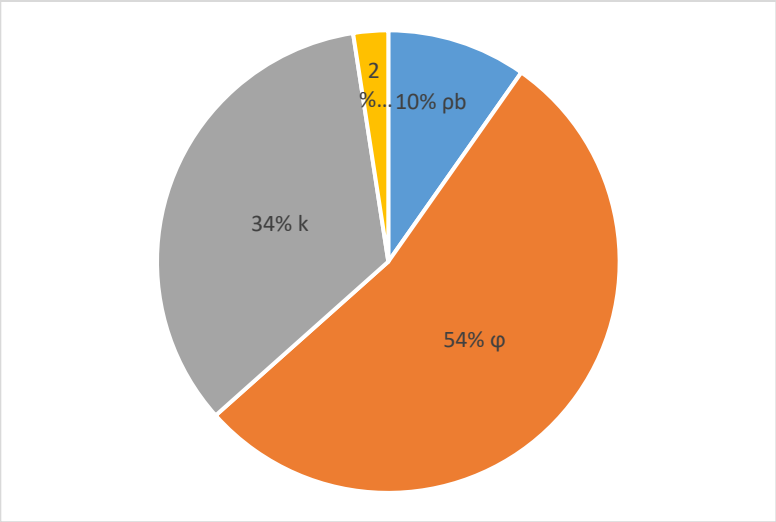


Figure28. Target Petrophysical Parameters Modeled in Real-Time ML Studies

Table 02. Summary of Research Studies on Real-Time Petrophysical Parameter Estimation Using Artificial Intelligence

Year	Input	Output	Algorithms	Reference
2025	DD/MGD	Bulk Density	CB, AB, LGBM, XGB	Boutaghane et al.
2024	DD	Permeability	FNS, GBR, ANNS	Hassaan et al.
2024	DD	Porosity, Permeability	RF, SVM, DT	Hassaan et al.
2024	Mud Gas Data	Porosity	ANN, RF, DT	Anifowose & Badawood
2024	DD Mud Gas Data	Porosity	RF, SVM	Anifowose et al.
2023	Drilling Data	Porosity	ANN, RF, SVR, XGB	Ouladmansour et al.
2023	Drilling Data	Porosity	GB	Alhowaish et al.
2023	GR and DD	Porosity, Bulk Density	MLR, KNN, RFR, SVM, ANN	Sharma et al.
2023	Mud Gas Data	Porosity	ANN, DT, RF	Anifowose et al.
2023	Mud Gas Data	Porosity	RF, ANN, DT	Anifowose et al.
2022	Mud Gas Data	Porosity	ELM, RFR, MLPNN, GRNN, SVR	Ameur-Zaimeche et al.
2022	DD	Porosity	XGB, K-means Clustering	Mohammadian et al.
2022	Logging and DD	Permeability	ANN, XGB	Qalandari et al.

Year	Input	Output	Algorithms	Reference
2022	DD	Porosity	LR, RF, ANN, XGB	Ouadi et al.
2022	DD	Porosity	ANN	Gamal & Elkatatny
2022	DD	Permeability	3D CNN	Elmorsy et al.
2022	DD	Porosity	GMDH, SVM, DT, PR	Kamali et al.
2022	Mud Gas Data	Porosity	ANN	Anifowose et al.
2022	DD	Porosity	CNN	Iklassov et al.
2022	DD	Bulk Density	FN, SVM, RF	Ahmed et al.
2022	DD	Bulk Density	ANN, ANFIS	Ahmed et al.
2022	DD	Porosity	-	Singh & Bhardwaj
2022	DD	Permeability	ANN	Nande & Patwardhan
2022	DD	Permeability	RF	Ezekiel et al.
2021	DD	Permeability	GA, ANN	Tian et al.
2021	DD	Porosity	RF, DT	Gamal et al.
2021	DD	Permeability	LS-SVM, PSO-NN	Farouk et al.
2021	DD	Permeability	LR, SVR, GB, CNN, HPDNN	Tembely et al.
2021	DD	Porosity	ANN	Al-Sabaa et al.
2021	DD	Permeability, Water Saturation	XGB, RF	Otchere et al.
2021	DD	Porosity	ANN	Clarke et al.
2020	DD	Porosity	-	Al-AbdulJabbar et al.
2020	DD	Porosity	-	Basu et al.
2020	DD	Permeability	GA, RF	Wang et al.
2020	DD	Porosity	MLSTM, RNN	Chen et al.
2019	DD	Permeability	DNNS	Arigbe et al.
2019	DD	Permeability	NN	Zolotukhin & Gayubov

Chapter 04

Materiel and Method

Introduction

In practice, permeability can be assessed through both direct physical measurements and advanced computational models. This chapter outlines the integrated workflow employed in this study, combining conventional methods—such as core analysis and well log interpretation—with unconventional, data-driven modeling based on machine learning.

The methodology begins with a thorough examination of available data sources, including real-time drilling parameters and mud gas measurements. These inputs are processed and analyzed through a structured pipeline involving data cleaning, statistical exploration, and machine learning implementation. A suite of algorithms, including Random Forest and Extreme Gradient Boosting, is selected based on their robustness and predictive accuracy in geoscientific applications.

In addition to model development, particular emphasis is placed on interpretability through the use of SHAP values, ensuring transparency in feature contributions. The entire workflow was executed using Python and a set of specialized libraries designed for data analysis and model training. This methodological framework not only bridges traditional and modern approaches but also enables scalable, real-time permeability estimation tailored to operational demands.

To anchor this modern approach within a reliable foundation, the study first incorporates conventional methods of permeability estimation. These established techniques not only provide essential reference values but also help contextualize and validate the results generated by machine learning. The following section outlines these traditional approaches, beginning with direct laboratory-based measurements.

1. Conventional Methods

1.1 Direct method

1.1.1 Core Permeability Estimation

Core analysis provides the most direct means of measuring rock-formation permeability under tightly controlled laboratory conditions and is often treated as the reference standard. However, laboratory core measurements reflect only the specific sample's conditions (pressure, temperature, saturation) and may not fully capture reservoir-scale flow behavior. When applied consistently over a given interval, though, core-derived permeabilities become invaluable for completion design—particularly for optimizing perforation phasing and vertical spacing.

1.1.1.1 Sampling scales

- Sidewall cores (< 1 in.): small chips taken through drilling or wireline tools; high risk of sampling bias in heterogeneous formations.
- Core plugs (1–1½ in.): extracted every ~6 in. from full-diameter core; standard for routine permeability testing but still limited in representative volume.
- Full-diameter cores (~6 in. segments): medium-scale samples that better average out small-scale heterogeneities.
- Whole-core (up to 2 ft): largest-scale samples, rarely available due to recovery difficulties, but offering the closest approach to in-situ rock volumes.

1.1.1.2 Absolute permeability

- Steady-state method: after cleaning and drying, air (or water) is flowed through the core at constant pressure differential; Darcy's law is used to compute permeability (Brace et al., 1968)
- Pulse-decay method: especially suited for tight rocks where true steady flow is impractical; a pressure pulse is applied and the subsequent decay curve is analyzed.
- Directional values: labs typically report three principal permeabilities—maximum horizontal (k_{max}), perpendicular horizontal (k_{90}), and vertical (k_v).
- Klinkenberg correction: applied to air-flow measurements to account for gas-slippage effects at low pressures, aligning them with liquid permeability.

1.1.1.3 Relative permeability

- Steady-state co-flow: core is fully saturated with the wetting phase (water), then oil and water are co-injected at controlled ratios; resulting pressures and flow rates yield phase-specific permeabilities as functions of saturation.
- Displacement (unsteady-state) method: core initially oil-saturated; water is injected to displace oil while monitoring pressure and effluent volumes, mapping relative-permeability curves.
- Pulse-decay adaptation: if pressure pulses remain below capillary pressure thresholds, the decay response can also be used to infer relative permeabilities.

By integrating measurements across these scales and techniques—and applying consistent lab protocols—core analysis delivers a robust foundation of permeability data. These core-derived values then serve both as direct inputs for completion strategies and as ground truth for calibrating well-log and machine-learning models.

1.2 Indirect Merhodes

1.2.1 Wireline-Log-Based Permeability Estimation

Wireline logging provides a continuous, in-situ means of estimating formation permeability, often supplementing or extrapolating beyond core measurements. Over the decades, several empirical and physics-based models have been developed to link wireline log responses with permeability, ranging from porosity-permeability correlations to complex interpretations of acoustic and magnetic responses.

1.2.1.1 Empirical Correlations from Porosity and Water Saturation

Early models aimed to relate permeability to porosity and other measurable properties. The Kozeny–Carman equation laid the foundation by linking permeability to porosity (ϕ) and the specific surface area of the grains (Ag):

$$k = \frac{\phi^3}{5A_g^2(1-\phi)^2} \quad (6)$$

However, Ag can only be determined from core samples, making this approach less practical. As a result, simpler empirical models were developed using log-derived parameters. A widely used example is the Timur equation, later improved by Coates and Dumanoir to account for irreducible water saturation

$$k = 100 \cdot \phi_e^2 \cdot \left(\frac{1 - S_{wi}}{S_{wi}} \right) \quad (7)$$

Here, ϕe is effective porosity. This model gives zero permeability when porosity or movable fluid approaches zero, making it more consistent with physical behavior.

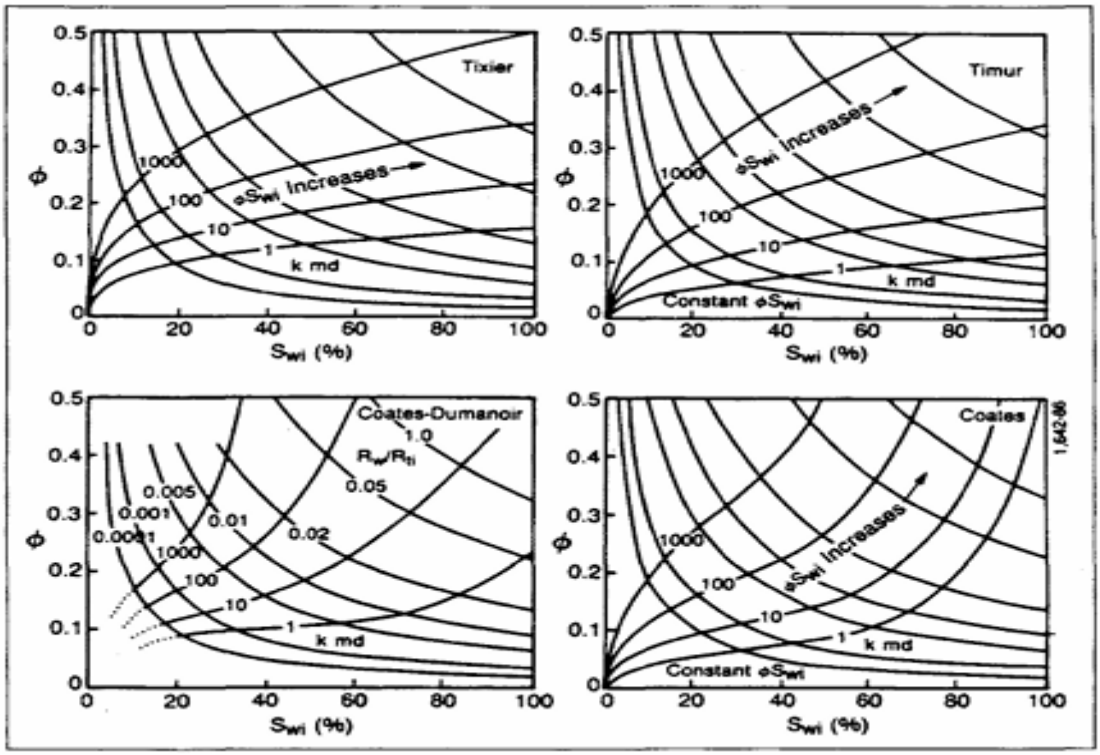


Figure 29. Charts for estimating permeability from porosity and water saturation

1.2.2 NML (Nuclear Magnetic Logging) Nuclear Magnetic Resonance Logging (NML)

NMR tools provide direct information about pore fluids and pore size. The key outputs are the free-fluid index (FFI) and T_1 or T_2 relaxation times, which relate to fluid mobility and pore geometry. Since permeability is proportional to the square of pore size, it can be related to T_1 (or T_2) through an empirical relationship:

$$k = 1.6 \times 10^{-9} T_2^{2.3} \phi^{4.3} \quad (8)$$

this method is especially useful in clean sandstones but may be less accurate in carbonates.

1.2.3 Geochemical Logging Tools (GLT)

GLT tools use borehole spectroscopy to measure elemental concentrations, which reflect changes in mineralogy and pore structure. By incorporating mineralogical data, a modified version of the Kozeny–Carman model can be applied

$$\log k = T_m + 3\log(\phi) - 2\log(1-\phi) + \sum B_i f_i \quad (9)$$

Where (Tm) is textural maturity, (Bi) are constants for specific minerals, and fi are their weight fractions. This method refines permeability estimation by considering mineral composition and rock texture.

1.2.4 Stoneley Wave Analysis from Sonic Logs

Stoneley waves are acoustic waves that travel along the borehole wall. In permeable zones, these waves are attenuated due to fluid flow into the formation. Though quantitative estimation is complex, Stoneley wave attenuation and dispersion can be used qualitatively to identify high-permeability or fractured zones.

1.2.5 Pressure Transient Analysis with Formation Testers (e.g., RFT, MDT)

Formation testers extract fluid samples and measure pressure changes over time. By analyzing drawdown and buildup responses, permeability can be estimated using pressure-transient theory. Superflow tests, which involve sustained fluid extraction, offer permeability estimates that reflect hydrocarbon productivity more accurately than initial formation tests.

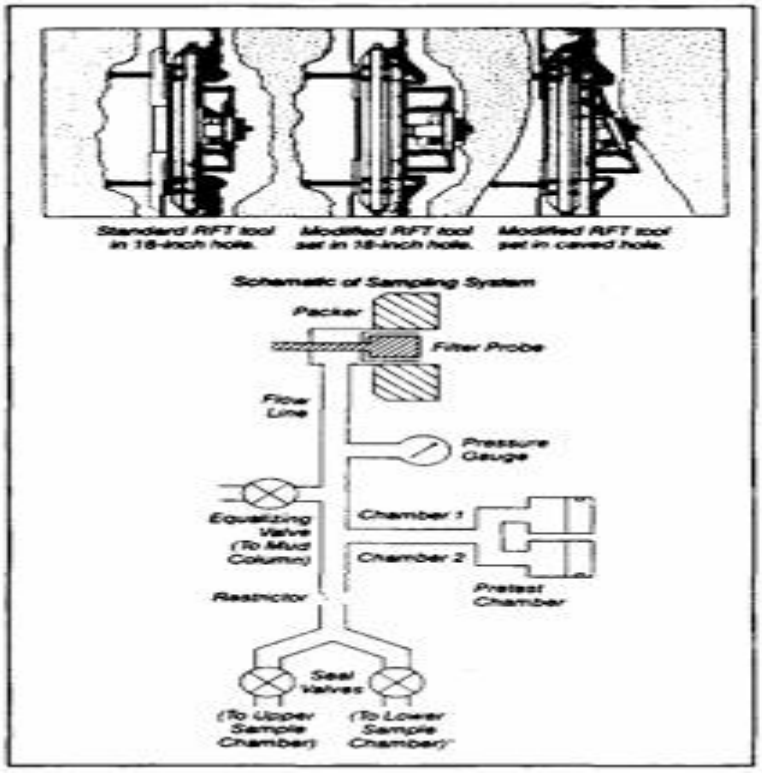


Figure 30. RFT tool setup and shematic of sampling system

1.2.6 Well Testing (Dynamic Field Methods)

Well testing is a dynamic method for estimating formation permeability and provides information over a much larger volume compared to core analysis or wireline logging. Well testing procedures can be categorized into three major types:

1.2.6.1 Short-Term Tests

These include:

- Drill Stem Testing (DST): Typically performed in open hole or cased sections with perforations. DST involves two drawdown and buildup cycles. Due to the short test duration, results may be influenced by wellbore storage and mud filtrate invasion.
- IMPULSE℠ Testing: Conducted immediately after perforation. Involves an instantaneous rate change followed by shut-in. The pressure response is analyzed to estimate permeability via type-curve or Homer plot methods.
- TRAP℠ Testing: Uses transient downhole pressure and rate to estimate formation permeability. TRAP reduces wellbore storage effects and provides accurate permeability values even with short test durations.

1.2.6.1 Conventional Testing

These include classic:

- Pressure Drawdown Tests: Involve flowing the well at a constant rate and monitoring the pressure drop over time.
- Pressure Buildup Tests: More common in practice. The well is shut-in after a period of steady production. The resulting pressure buildup is analyzed using superposition time functions to estimate permeability. These tests provide a good approximation of near-wellbore reservoir permeability.

1.2.6.2 Advanced Testing Techniques

Advanced methods are used for more complex reservoir assessments:

- Layered Reservoir Testing: Applies transients to individual layers and analyzes pressure/flow response using production logging tools.
- Vertical Interference Testing: Induces a transient in one zone and measures the response in another, evaluating vertical permeability and communication.
- Multiwell Interference Testing: A transient is applied in one well and pressure is recorded in another. This technique evaluates horizontal reservoir continuity and provides average permeability-thickness (kh) estimates.

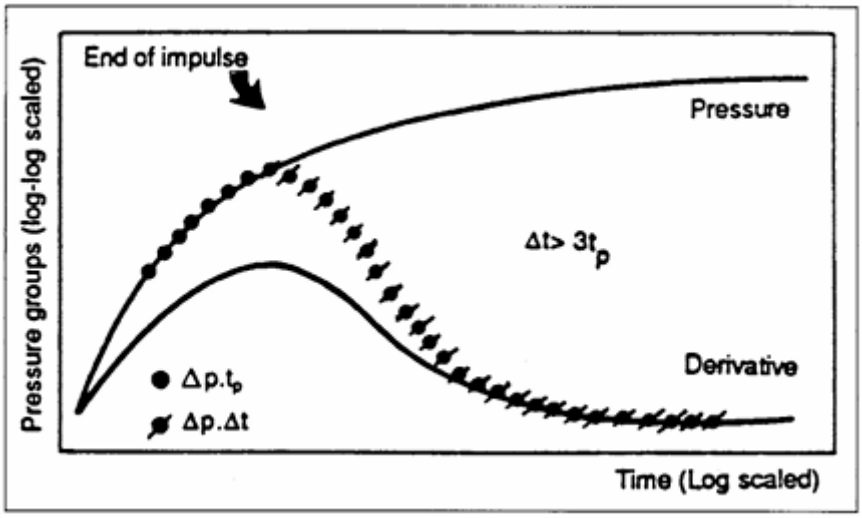


Figure 31. Typical Impulse test plot

2.Unconventional Method

Unconventional methods leverage data-driven techniques and real-time measurements to estimate permeability, moving beyond traditional laboratory and logging-based approaches (Mellal et al., 2023) . These methods harness the power of machine learning algorithms to model complex, nonlinear relationships between drilling parameters, mud gas data, and formation properties

In order to predict Permeability values from the drilling parameters , the following steps have been followed:

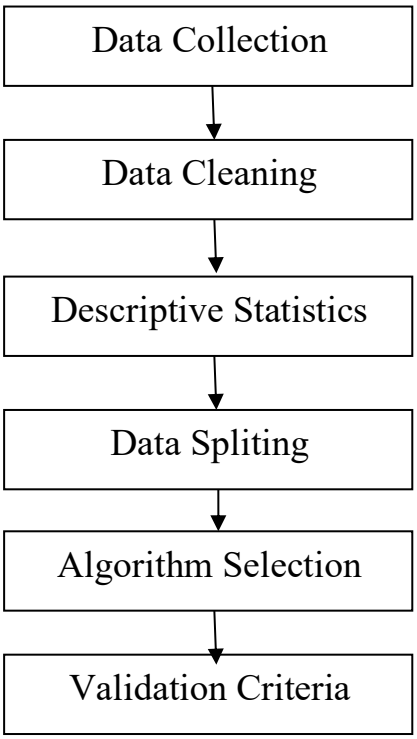


Figure 32. Methodological framework

2.1 Data Collection

The data used for this analysis were obtained from Sonatrach and consist of real-time measurements acquired during drilling operations. These data represent continuous recordings from the field, capturing key operational information necessary for further processing and modeling. The dataset forms the foundation for estimating permeability using data-driven techniques.

2.2 Data Cleaning

Before conducting any analysis, the raw data underwent a comprehensive cleaning process to ensure consistency, accuracy, and reliability. This step involved handling missing values, removing duplicates, and correcting any irregularities or anomalies within the dataset. The objective was to prepare a high-quality dataset suitable for modeling and interpretation, minimizing potential errors during the machine learning workflow.

2.3 Descriptive Statistics

2.3.1 Univariate statistics: Univariate data refers to a type of data in which each observation or data point corresponds to a single variable. In other words, it involves the measurement or observation of a single characteristic or attribute for each individual or item in the dataset. Analyzing univariate data is the simplest form of analysis in statistics.
Key points in Univariate analysi*:*

- No Relationships: Univariate analysis focuses solely on describing and summarizing the distribution of the single variable. It does not explore relationships between variables or attempt to identify causes.

- Descriptive Statistics: such as measures of central tendency (mean, median, mode) and measures of dispersion (range, standard deviation), are commonly used in the analysis of univariate data.
- Visualization: Histograms, box plots, and other graphical representations are often used to visually represent the distribution of the single variable.

2.3.2 Bivariate data: Bivariate data involves two different variables, and the analysis of this type of data focuses on understanding the relationship or association between these two variables

Key points in Bivariate analysis:

- Relationship Analysis: The primary goal of analyzing bivariate data is to understand the relationship between the two variables. This relationship could be positive (both variables increase together), negative (one variable increases while the other decreases), or show no clear pattern.
- Scatterplots: A common visualization tool for bivariate data is a scatterplot, where each data point represents a pair of values for the two variables. Scatterplots help visualize patterns and trends in the data.
- Correlation Coefficient: A quantitative measure called the correlation coefficient is often used to quantify the strength and direction of the linear relationship between two variables. The correlation coefficient ranges from -1 to 1. (Kotz, S.; et al., 2006)

2.3.3 Multivariate data: Multivariate data refers to datasets where each observation or sample point consists of multiple variables or features. These variables can represent different aspects, characteristics, or measurements related to the observed phenomenon. When dealing with three or more variables, the data is specifically categorized as multivariate.

Key points in Multivariate analysis:

- Analysis Techniques:The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, principal component analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).
- Goals of Analysis: The choice of analysis technique depends on the specific goals of the study. For example, researchers may be interested in predicting one variable based on others, identifying underlying factors that explain patterns, or comparing group means across multiple variables.
- Interpretation**:** Multivariate analysis allows for a more nuanced interpretation of complex relationships within the data. It helps uncover patterns that may not be apparent when examining variables individually.

2.4 Data Spliting

To ensure that a machine learning model performs well on unseen data and avoids common pitfalls such as underfitting or overfitting, it is essential to split the available dataset into distinct subsets: training, validation, and test datasets. Brownlee, J. (2017, July)

Underfitting occurs when a model is too simplistic or not trained adequately, resulting in poor performance on both training and unseen data. Overfitting, on the other hand, arises when a model is trained excessively on the training data, including its noise and outliers, leading to high training accuracy but poor generalization to new data. To mitigate these issues, dataset partitioning is a standard practice.

- **Training data**: A subset of observations used for the model training process. This dataset is used by the algorithm to learn the parameters of the machine learning model.
- **Validation data**: A subset of observations, which is used for evaluating the prediction performance of the model during the training process. The trend observed in prediction errors using the validation data can help in identifying the number of iterations for adequate training. In a typical model training, the prediction error should keep on decreasing for both the training and validation dataset through the iterative training process. However, if we observe a point where prediction error continues to decrease for the training data but starts to increase for the validation data, training should be stopped. The point from where the validation error begins to increase indicates the beginning of overfitting
- **Test data:** This is the subset of data that is only used to assess the performance of a fully trained model. The prediction accuracy on the test data is an indicator of the model's performance on unseen data encountered in a real-world scenario.

Typical data splitting strategies include a simple 80% training and 20% test division. More robust setups use a 70% training, 15% validation, and 15% testing split to ensure both model tuning and final evaluation are unbiased and comprehensive. This structured division of data ensures the model is neither undertrained nor excessively fine-tuned, supporting balanced performance across different data subsets and ultimately promoting generalization

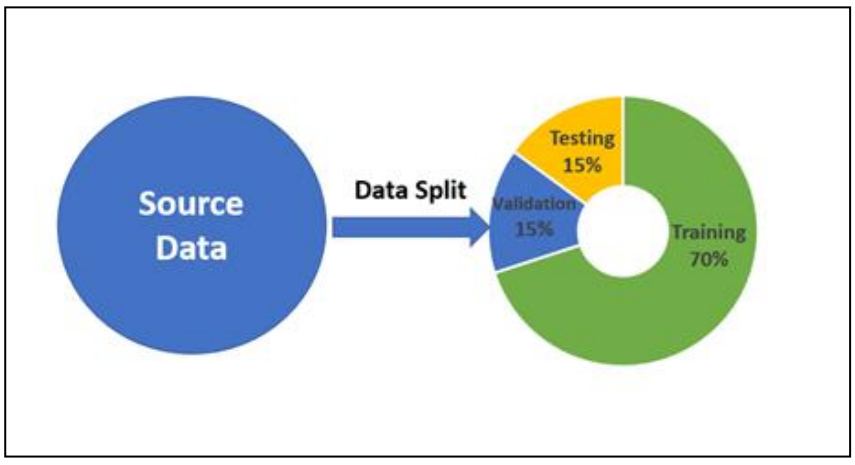


Figure 33. Typical Machine Learning Data Split: Training, Validation, and Testing Sets

2.5 Algorithm selection

The selection of a machine learning algorithm is primarily influenced by the intrinsic characteristics of the dataset, including its size, structure, and quality, as these factors determine the algorithm's suitability and robustness. The nature of the problem—whether classification, regression, or clustering—also dictates the choice of algorithms, as certain methods are specifically designed for particular problem types. For large datasets, computationally efficient and scalable algorithms, such as Random Forests or Gradient Boosting Machines, are generally

preferred due to their capacity to model complex patterns without excessive computational cost. In contrast, for smaller datasets, simpler models like Logistic Regression or Support Vector Machines are more appropriate because they are less prone to overfitting and demand fewer computational resources while maintaining reliable performance.

2.6 Validation criteria

To effectively assess the performance of machine learning models and identify the most suitable one, it is essential to implement a reliable validation strategy alongside carefully chosen evaluation metrics that align with the specific objectives of the problem.

- ✓ **The correlation coefficient (R):** measures the linear relationship between predicted and observed values in petrophysical modeling, with values near +1 indicating strong positive correlation. Interpretative thresholds (e.g., 0.40–0.69 moderate, 0.70–0.89 strong) vary and are context-dependent (Rodgers & Nicewander, 1988). R should be evaluated with caution, considering data range, confidence intervals, and practical relevance, as statistical significance does not always imply meaningfulness. This balanced approach is crucial for reliable model assessment.

$$R = \left[\frac{\frac{1}{N}\sum_{i=1}^{N}(XI)\cdot(\overline{XI})}{\sqrt{(XI-\overline{XI})^2}\cdot\sqrt{(YI-\overline{YI})^2}}\right] ,(-1 < R < +1) \qquad (10)$$

- ✓ **The coefficient of determination (R^2)**: quantifies the proportion of variance in the dependent variable explained by the regression model, with values ranging from 0 (no explanatory power) to 1 (perfect fit). Negative R^2 indicates poorer performance than a mean-based model. Since R^2 depends on the dataset, it should not be compared across different datasets.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (11)$$

- ✓ **Root mean square error: i**s a common metric to evaluate regression models. It measures how far predictions are from actual values using squared differences. The formula is

$$RMSE = \sqrt{\frac{1}{N}(XI - YI)^2}, (0 \leq RMSE < +\infty) \qquad (12)$$

MAE (Mean Absolute Error): on the other hand, corresponds to the Manhattan **norm**, also known as the $\ell 1$ norm. This measure calculates the sum of the absolute differences between predictions and actual values. Unlike RMSE, MAE does not square the errors, making it less sensitive to large deviations or outliers. This results in a more stable error measure that treats all errors equally

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|YI - XI|, (0 \leq MAE < +\infty) \quad (13)$$

- ✓ **The Nash-Sutcliffe Efficiency (NSE)** : measures how well model predictions match observed data relative to the mean. An NSE of 1 indicates perfect prediction, while values ≤ 0 imply poor performance.

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N}[XI - YI]^2}{\sum_{i=1}^{N}[XI - \overline{YI}]^2}\right], \ (-\infty < NSE \leq 1) \quad (14)$$

2.7 Model Interpretability Tools

Interpreting machine learning models is essential for understanding how predictions are made and building trust in their results. This section introduces key interpretability methods—SHAP and feature importance—that help explain the influence of individual features on model outcomes, applicable across various types of machine learning models.

2.7.1 SHAPLY Additive: is a model-agnostic interpretability method based on cooperative game theory. It assigns each feature an importance value representing its contribution to a specific prediction. By approximating Shapley values, SHAP provides consistent and locally accurate explanations across various machine learning models.

2.7.2 Feature **importance:** refers to techniques that assign a score to input features based on how useful they are in predicting a target variable. It helps identify key drivers of model predictions and improves interpretability. Common methods include permutation importance, tree-based scores, and model-agnostic approaches like correlation analysis.

3. Softwares

Machine learning software refers to specialized programs and frameworks that support the automation of analytical model building. These tools provide the infrastructure for feeding data into algorithms, adjusting model parameters, and evaluating outputs. By enabling systems to improve performance through experience, machine learning software plays a crucial role in applications such as fraud detection, recommendation systems, and real-time forecasting across industries.

3.1 Python: is a widely used, high-level programming language known for its simplicity, readability, and versatility. It supports multiple programming paradigms, including object-oriented and procedural programming. Python is platform-independent, open-source, and has a large ecosystem of libraries, making it especially popular for machine learning, data science, and automation tasks. Its intuitive syntax and dynamic typing make it easy to learn and use, even for beginners.

3.2 Excel: Microsoft Excel is a powerful spreadsheet application used to enter, organize, and analyze data in a tabular format. It offers a grid of rows and columns where users can perform calculations, create charts, and manage data efficiently. Excel is widely utilized across various fields for tasks ranging from simple budgeting to complex data analysis.

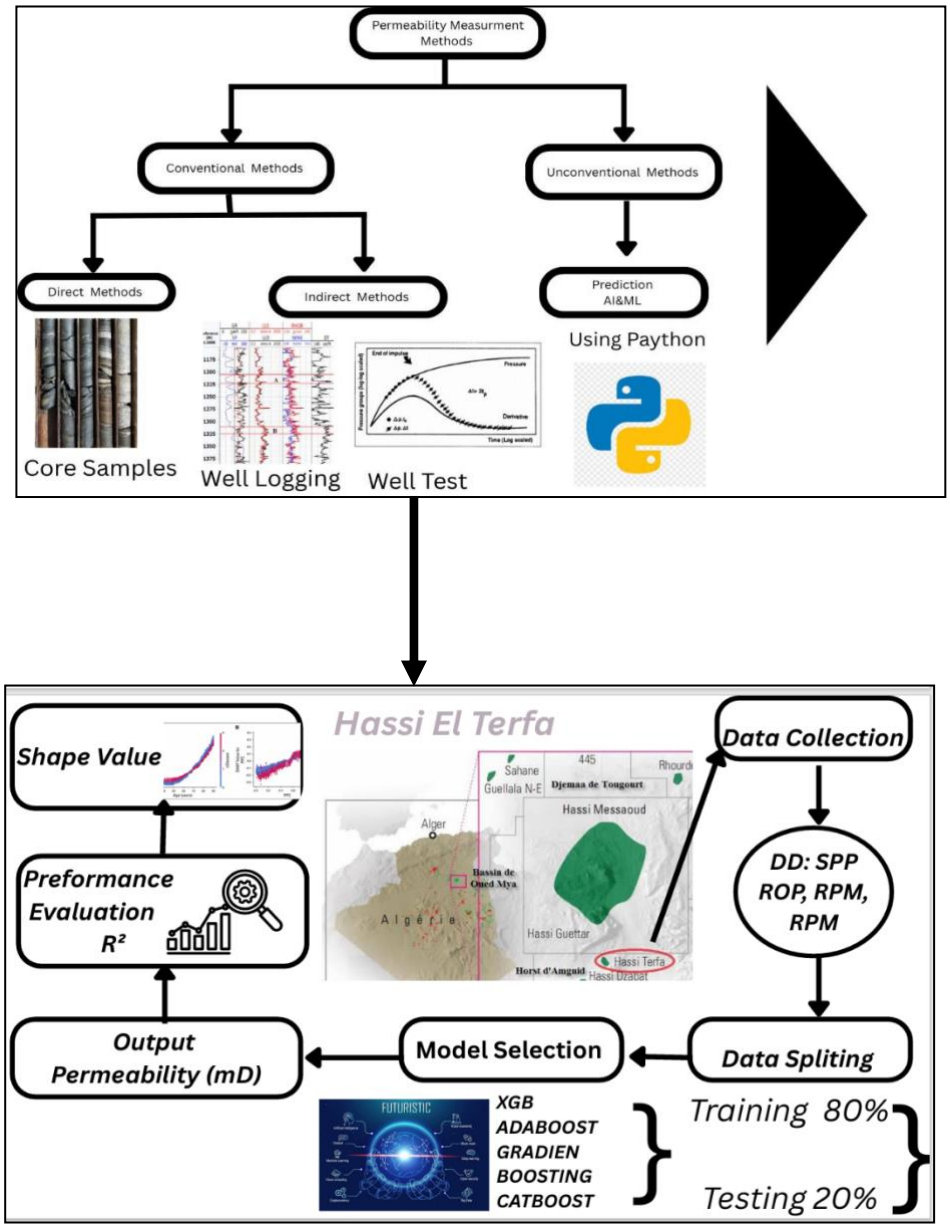


Figure 34. Flowchart summary of Permeability Prediction method

Chapter 05

Results & Discussions

Introduction

This chapter presents the application of various machine learning algorithms to predict a key petrophysical parameter from real-time drilling data. The aim is not only to assess the predictive performance of different models but also to interpret their decisions using explainable AI tools such as SHAP (SHapley Additive exPlanations). The models analyzed include XGBoost Regressor, CatBoost Regressor, AdaBoost Regressor, and Gradient Boosting Regressor. This approach enables a deeper understanding of the contribution of each input feature toward the final prediction.

1.Model Deployment Setup

To evaluate model performance, the original dataset of 183 samples was randomly split into training and testing subsets: 146 samples (80%) for training and 37 samples (20%) for testing. This split ensures that model evaluation reflects the ability to generalize to unseen data.

Each model was trained to predict permeability based on six real-time drilling parameters:

- Weight on Bit (WOB),
- Rotations per Minute (RPM),
- Torque,
- Flow Rate (Q),
- Standpipe Pressure (SPP),
- Rate of Penetration (ROP).

During deployment, predictions were generated exclusively for the test set. These predicted permeability values were then compared to actual measurements to assess performance using quantitative metrics such as RMSE, R^2, MAE, and Nash-Sutcliffe Efficiency (NSE), alongside graphical analyses including scatter and residual plots.

This setup establishes a clear framework for interpreting the subsequent results.

2. Results and Discussion

Before modeling, a comprehensive statistical analysis was performed to understand the nature of the input features and the target (permeability). The descriptive statistics are summarized as follows

2.1 Univariate Statistics

The data shows considerable variation, especially in Torque and Permeability. Torque ranges widely, from 0 up to 6865, indicating some extreme values or outliers. Permeability has a median of 0.526, which is much lower than its mean (31.28), suggesting a skewed distribution with many low values and few high values. Other parameters such as WOB and RPM show more consistent distributions around their means. These characteristics will be important when interpreting model performance and results.

Table 03. Univariate analysis

	WOB	RPM	Torque	Q	SPP	ROP	PERMEABILITY (mD)
count	182.000000	182.000000	182.000000	182.000000	182.000000	182.000000	182.000000
mean	4.992917	67.214409	1474.179155	1194.453670	1683.994390	32.251632	31.276811
std	2.681390	22.569920	2525.359254	407.292055	772.825412	20.673555	56.360111
min	1.063613	0.000000	0.000000	894.726991	1085.737179	6.050000	0.010738
25%	2.955435	58.373755	200.845948	917.507924	1161.286600	10.317000	0.050275
50%	4.524181	59.101337	219.130306	946.382844	1194.139194	34.658500	0.526074
75%	7.036177	90.913411	275.111309	1787.000000	2805.750000	41.298000	34.683520
max	10.792598	99.944284	6865.000000	1971.000000	3099.000000	140.542000	274.402200

2.2 Bivariate Statistics Pearson correlation analysis was performed to assess the linear relationships between permeability and the drilling parameters. As illustrated in the correlation heatmap (Figure 35), permeability is negatively correlated with all input features, suggesting an inverse relationship between drilling intensity and reservoir quality.

The strongest negative correlations were observed between permeability and:

- Weight on Bit (r = -0.45),
- Rate of Penetration (r = -0.44),
- Rotations per Minute (r = -0.34).

Weaker inverse correlations were noted with flow rate (r = -0.20), standpipe pressure (r = -0.17), and torque (r = -0.15). These findings suggest that increased drilling force and speed are generally associated with lower formation permeability, possibly due to compaction or drilling-induced damage.

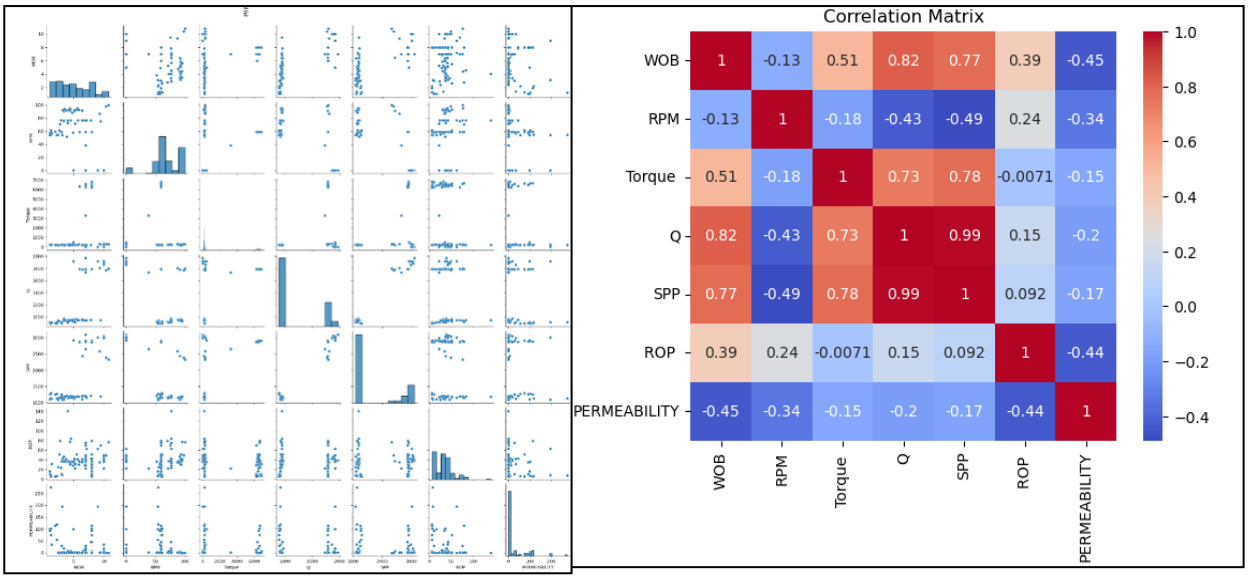


Figure 35. Pair plot & Correlation Heatmap of Drilling Parameters

2. 3 Model Performance Evaluation

This section presents the evaluation results of four ensemble learning models—XGBoost Regressor**,** CatBoost Regressor**,** AdaBoost Regressor**,** and Gradient Boosting Regressor—developed for real-time formation permeability prediction. The models were assessed based on their performance on the training dataset and subsequently validated on a separate testing dataset. The evaluation metrics included the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and Nash–Sutcliffe efficiency (NSE)

2.3.1 Training Performance:The results on the training data indicate that most models were able to capture the underlying relationships within the dataset effectively. Table (04) summarizes the training performance.

Table 4. Initial Training Performance Metrics for Permeability Prediction Models

Model	R^2	RMSE	MAE	NSE
XGBoost Regressor	1.0000	0.0009	0.0006	1.0000
CatBoost Regressor	0.9996	1.0272	0.6849	0.9996
AdaBoost Regressor	0.8209	22.9340	18.0936	0.8209
Gradient Boosting Regressor	0.9955	3.6431	2.2679	0.9955

The XGBR achieved a perfect fit with an R^2 of 1.000, followed closely by CBR and GBR, both showing excellent training performance. The ABR, while slightly less accurate, still delivered strong results with an R^2 above 0.82. These high values indicate that all models effectively captured the underlying patterns in the training data.

2.3.2 Testing Performance: To evaluate how well the models generalized to unseen data, their performance was further tested using a holdout dataset. The results are shown in Table (05)

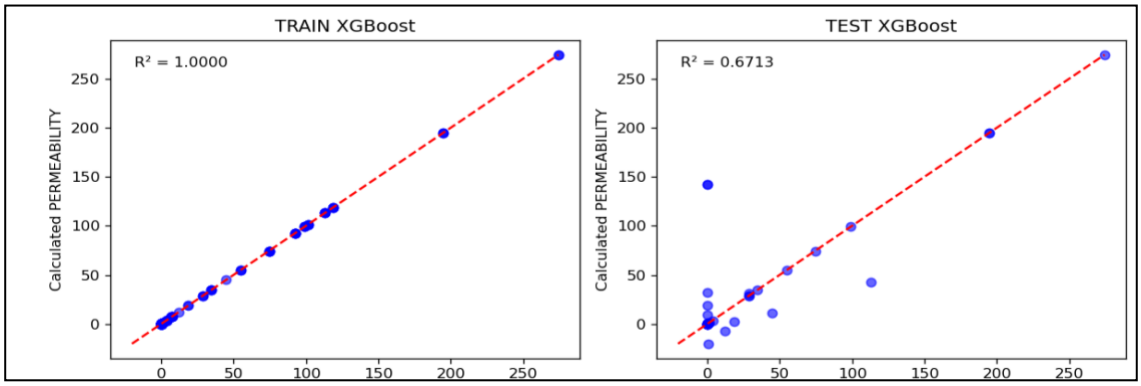
Table 05. Initial Testing Performance Metrics for Permeability Prediction Models.

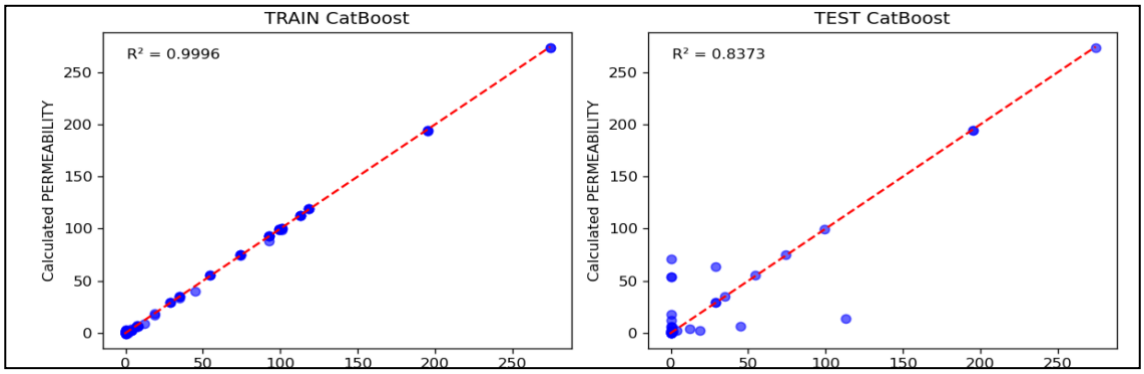
Model	R^2	RMSE	MAE	NSE
XGBoost Regressor	0.671269	36.398578	13.767598	0.671269
CatBoost Regressor	0.837283	25.608287	11.863584	0.837283
AdaBoost Regressor	0.746010	31.994254	24.383186	0.746010
Gradient Boosting Regressor	0.578498	41.215806	17.504057	0.578498

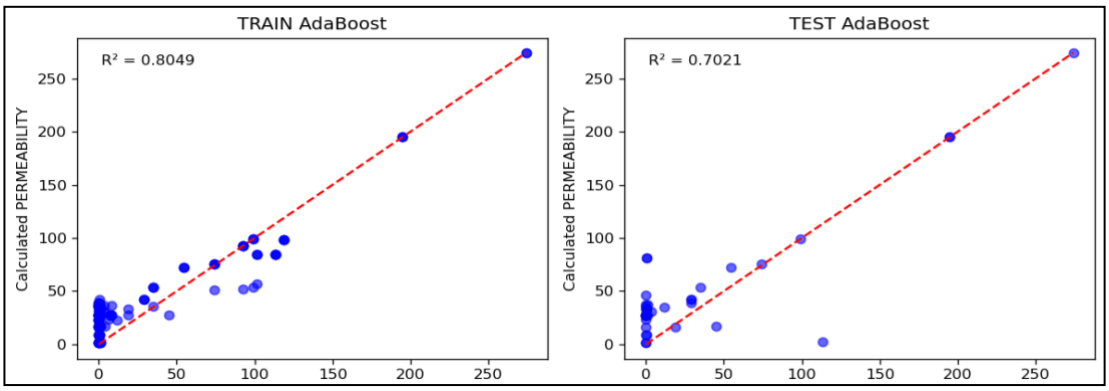
The testing results highlight a clear contrast in the generalization ability of the models. While all models performed well during training, their predictive accuracy on unseen data varied significantly

The CBR emerged as the most robust model, achieving the highest R^2 (0.837) and the lowest RMSE (25.61) among all tested models. This indicates strong generalization and minimal overfitting. In contrast, XGBR, despite its perfect training score, showed a noticeable performance drop on the test set (R^2 = 0.671), suggesting potential overfitting. The ABR

delivered moderate performance on the testing dataset, while GBR recorded the lowest predictive accuracy ($R^2 = 0.578$), indicating limited generalization capability in this context.







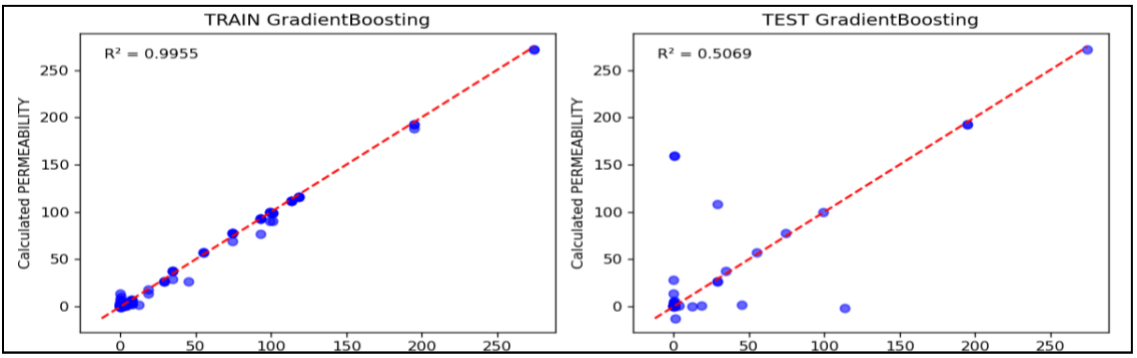


Figure 36. Training vs. Testing Performance of Regression Models

2.4 Best Feature Subset Performance

To improve predictive accuracy and reduce complexity, feature selection was applied to each ensemble model. The optimal results are shown in Table 06.

XGBR, CBR, and ABR all performed best using Torque, Flow Rate (Q), and Rate of Penetration (ROP). XGBR achieved the highest accuracy with an R^2 of 0.9655 and RMSE of 11.79.

For GBR, the best features were RPM, Q, and ROP, resulting in an R^2 of 0.9048.

These findings highlight the value of targeted feature selection in improving model and efficiency.

Table 06. Best performance metrics with optimized feature subsets for each model

Model	Best Features	R^2	RMSE	NSE
XGBoost Regressor	TORQUE, Q ROP	0,965518	11,788529	0,965518
CatBoost Regressor	TORQUE, Q ROP	0,924132	17,486133	0,924132
AdaBoost Regressor	TORQUE, Q ROP	0,889369	21,115503	7,4419149
Gradient Boosting Regressor	RPM, Q ROP	0,904771	19,590671	8,597870

2.5 Feature Importance and SHAP Value Analysis

The selection of input features for the ensemble machine learning models was informed by comprehensive feature importance and SHAP (SHapley Additive exPlanations) value analyses, which provided both quantitative and interpretative insights into the relative contribution of each variable.

Traditional feature importance metrics identified variables such as Weight on Bit (WOB), Flow Rate (Q), Rate of Penetration (ROP), and Torque as the most influential predictors, indicating their substantial impact on reducing model error and enhancing prediction accuracy.

Complementing this, the SHAP value analysis offered a detailed, sample-level explanation of the direction and magnitude of each feature's effect on permeability predictions. Specifically, the SHAP summary plots revealed that higher values of WOB and Q consistently increased the predicted permeability, while ROP exhibited a nuanced but generally positive relationship with the target variable. This dual approach of feature ranking and localized interpretability not only substantiates the selection of these features as the optimal subset but also aligns with the physical understanding of drilling dynamics influencing formation permeability.

Consequently, the reduced feature set derived from these analyses contributed to improved model performance by minimizing noise and enhancing generalization, thereby reinforcing the validity of the chosen predictors within both a statistical and domain-specific context

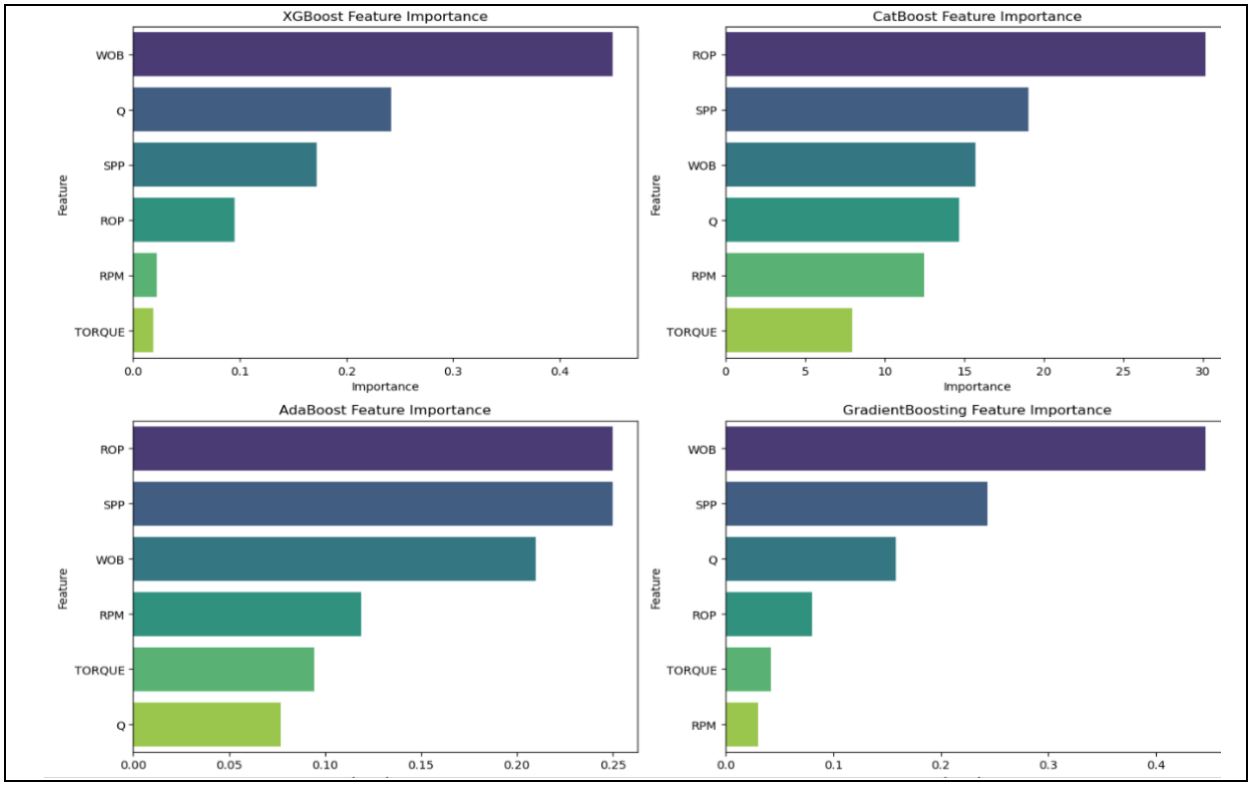
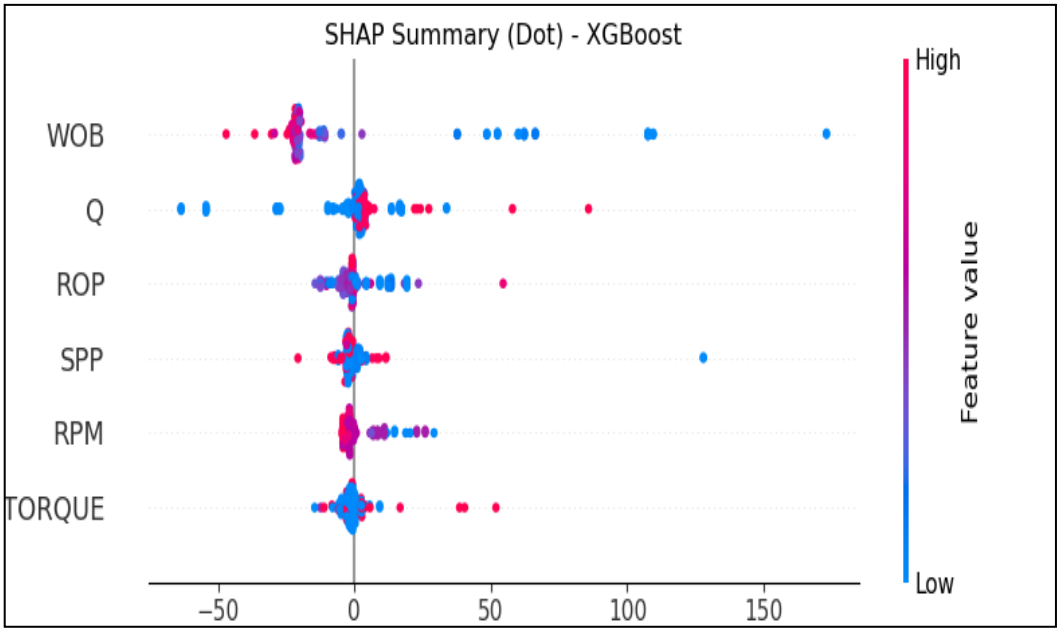


Figure 37. Feature Importance Rankings for Permeability Prediction Models



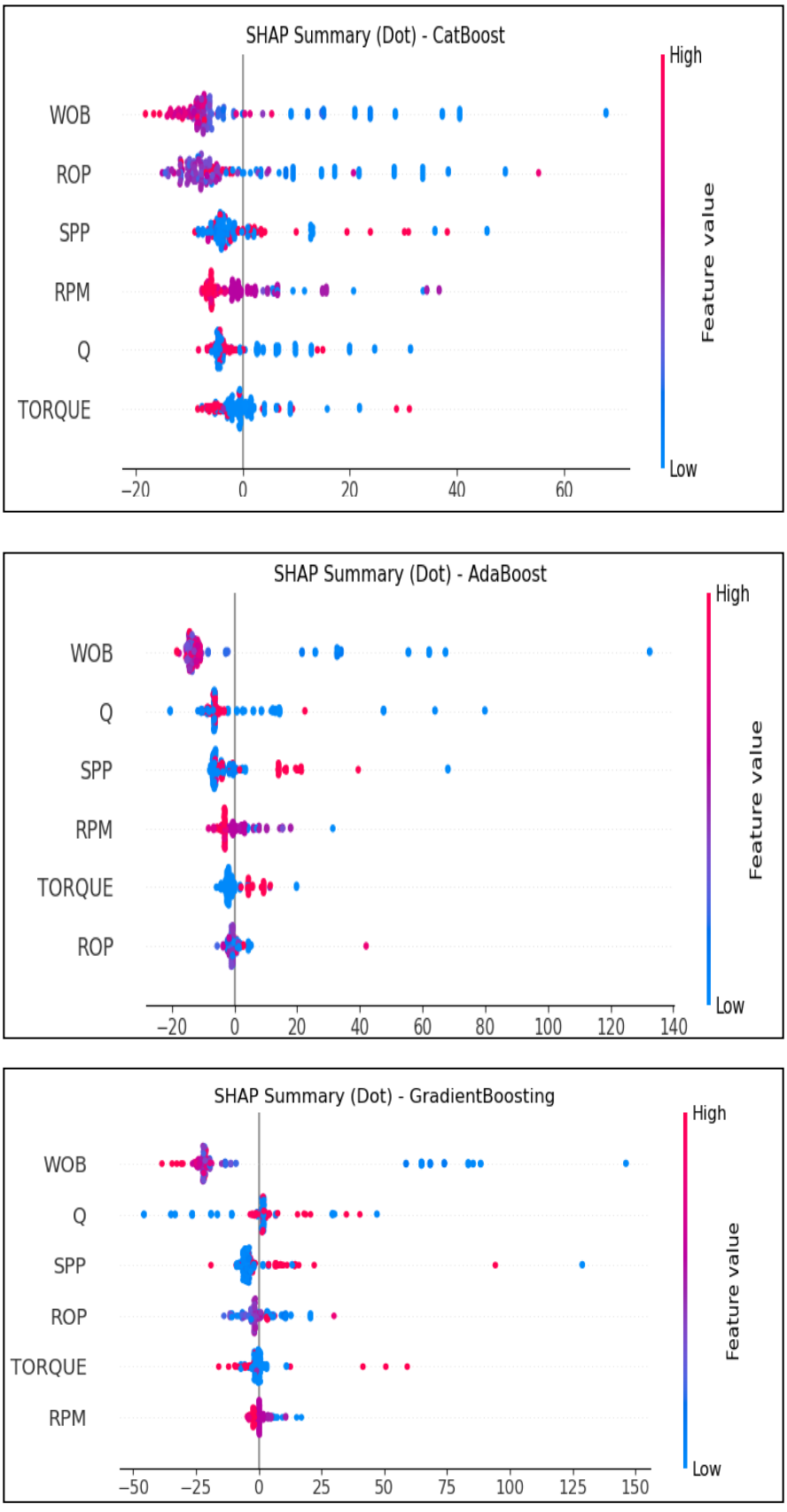


Figure 38. SHAP Summary Plots

Recommendations

In light of the findings from this study on real-time permeability prediction using surface drilling data and machine learning models, the following recommendations are proposed to guide future research and industry application:

1. **Operational Implementation of XGBoost Models**

   The XGBoost Regressor, which achieved the highest test R^2 score of 0.9655, demonstrated excellent performance in predicting formation permeability in real time. It is therefore recommended for integration into drilling workflows, particularly in regions with geological similarities to the Hassi Tarfa Field. Its compatibility with SHAP-based interpretability also makes it suitable for real-time operational environments where transparency is essential.

2. **Expansion of Dataset Size and Scope**

   Although this study yielded promising results using a dataset of 183 samples, the robustness and generalization of the model can be further improved by incorporating a larger and more diverse dataset. Future work should include data from multiple wells, varying lithologies, and different depositional environments.

3. **Development of Real-Time Monitoring Tools**

   The deployment of the XGBoost model in a user-friendly, real-time monitoring interface would provide field engineers with actionable insights during drilling operations. Incorporating SHAP visualizations into these tools can help operators understand the influence of each input feature on model predictions.

4. **Adoption of Physics-Informed Machine Learning (PIML)**

   Incorporating domain knowledge through Physics-Informed Machine Learning (PIML) frameworks can improve model reliability, particularly in extrapolative scenarios where purely data-driven models may be limited. This approach ensures that predictions remain consistent with geological and physical principles.

5. **Extension to Multi-Target Prediction**

   The methodology adopted in this study could be extended to simultaneously predict other petrophysical properties such as porosity, bulk density, and fluid saturation. A multi-target prediction framework could support more comprehensive reservoir characterization and better inform decision-making.

General Conclusion

In this thesis, we successfully demonstrated the potential of explainable machine learning models to predict formation permeability in real-time using surface drilling data from the Hassi Tarfa Field. By systematically applying and evaluating ensemble learning methods—namely XGBoost, CatBoost, AdaBoost, and Gradient Boosting—we highlighted how data-driven techniques can augment traditional geological understanding.

The application of SHAP analysis further reinforced the interpretability of the models, identifying Torque, Flow Rate, and Rate of Penetration as key contributors to permeability prediction. These findings confirm the significance of integrating domain knowledge with advanced analytics.

Our research not only contributes to the growing body of knowledge on real-time reservoir characterization but also sets a foundation for operational deployment in similar geological settings. Future work should explore larger datasets, integration with downhole measurements, and hybrid physics-informed models to further enhance reliability.

Ultimately, this thesis underscores that machine learning, when applied thoughtfully and interpretably, is not a replacement but a powerful complement to geological expertise in modern reservoir evaluation.

Bibliography

- Al-AbdulJabbar, A., Al-Azani, K., & Elkatatny, S. (2020). Estimation of reservoir porosity from drilling parameters using artificial neural networks. *Petrophysics*, *61*(03), 318–330.
- Alhowaish, H. A., Mezghani, M. M., & Shakriov, A. (2023, October). Predicting Porosity from Drilling Data Using Machine Learning–Challenges and Solutions. In *Abu Dhabi International Petroleum Exhibition & Conference*.
- Ameur Zaimeche, O., Zeddouri, A., Kouadria, T., Kechiched, R., & Belaksier, M. S. (2014, March). Use of Cluster Analysis method in log's data processing: prediction and rebuilding of lithologic facies. In *International Conference on Environmental Science and Geoscience (ESG'14) Venice, Italy* (pp. 98-101).
- Ameur-Zaimeche, O., Kechiched, R., Bouhafs, R., Mammeri, A., Hamdat, A., Zeddouri, A. (2022). Volume of Clay Estimation Using Artificial Neural Network Case Study: Berkine Basin Southern Algeria. In: Meghraoui, M., et al. *Advances in Geophysics, Tectonics and Petroleum Geosciences. CAJG 2019. Advances in Science, Technology & Innovation*. Springer, Cham. https://doi.org/10.1007/978-3-030-73026-0_81
- Ameur-Zaimeche, Ouafi., Kechiched, Rabah., Aouame Charafeddine. 2021. "Rate of penetration prediction in drilling wells from the Hassi Messaoud oil field (SE Algeria): Use of artificial intelligence techniques and environmental implications". In book: *Computers in Earth and Environmental Sciences Artificial Intelligence and Advanced Technologies in Hazards and Risk Management* Publisher: Elsevier: https://doi.org/10.1016/B978-0-323-89861-4.00032-4
- Ameur-Zaimeche, Ouafi., Kechiched, Rabah., Heddam, Salim., David A. Wood. 2022. Real-time porosity prediction using gas-while-drilling data and machine learning with reservoir associated gas: Case study for Hassi Messaoud field, Algeria. *Marine and Petroleum Geology*. Publisher: Elsevier. https://doi.org/10.1016/j.marpetgeo.2022.105631
- Ameur-Zaimeche, Ouafi., Zeddouri, Aziez., Heddam, Salim., Kechiched, Rabah., 2020. Lithofacies prediction in non-cored wells from the Sif Fatima oil field (Berkine basin, southern Algeria): A comparative study of multilayer perceptron neural network and cluster analysis-based approaches, *Journal of African Earth Sciences*, *Volume 166*, 2020, 103826, ISSN 1464-343X, https://doi.org/10.1016/j.jafrearsci.2020.103826
- Anifowose, F. A., Mezghani, M. M., Torlov, V., & Badawood, S. M. (2024, April). Predicting Rock Properties from Formation Fluid Measurements: Examples, Challenges, and Future Possibilities. In *SPE EOR Conference at Oil and Gas West Asia* (p. D021S021R002). SPE.
- Anifowose, F., Mezghani, M., Badawood, S., & Ismail, J. (2022, February). A first attempt to predict reservoir porosity from advanced mud gas data. In *International Petroleum Technology Conference* (p. D021S042R001). IPTC.
- Arigbe, O. D., Oyeneyin, M. B., Arana, I., & Ghazi, M. D. (2019). Real-time relative permeability prediction using deep learning. *Journal of Petroleum Exploration and Production Technology*, *9*, 1271–1284.
- Bantan, R. A., Ali, A., Naeem, S., Jamal, F., Elgarhy, M., & Chesneau, C. (2020). Discrimination of sunflower seeds using multispectral and texture dataset in combination with region selection and supervised classification methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(11), 113142.

Bibliography

- Boumelit, K., & Lefilef, M. (2019). *Contexte géologique et étude géophysique, géomécanique des roches réservoirs du champ de Hassi Tarfa (Hassi Messaoud, Wilaya de Ouargla)*. Master's Thesis, Université Mohammed Seddik Ben Yahia, Jijel.
- Boutaghane, A., Ameur-Zaimeche, O., Heddam, S., Kechiched, R., Tahar-Belkacem, N., Ouladmansour, A., ... & Wood, D. A. (2025). Enhancing formation bulk density prediction while drilling using mud logging data and interpretable boosting machine learning. *Earth Science Informatics*, *18*(1), 172. https://doi.org/10.1007/s12145-024-01642-7
- Brace, W. F., Walsh, J. B., & Francos, W. T. (1968). Permeability of Granite Under Pressure. *Journal of Geophysical Research*, *69*, 259–273.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.
- Brownlee, J. (2017). What is the Difference Between Test and Validation Datasets? *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/difference-test-validation-datasets/
- Chen, H., & Mohaghegh, S. (2018). Permeability estimation using intelligent data-driven models. *Journal of Natural Gas Science and Engineering*, *57*, 301–313.
- Chen, W., Yang, L., Zha, B., Zhang, M., & Chen, Y. (2020). Deep learning reservoir porosity prediction based on multilayer long short-term memory network. *Geophysics*, *85*(4), WA213–WA225.
- Dandekar, A. Y. (2006). *Petroleum Reservoir Rock and Fluid Properties*. CRC Press.
- Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.
- Ellis, D. V., & Singer, J. M. (2007). *Well Logging for Earth Scientists*. Springer Science & Business Media.
- Farouk, S., Sen, S., Ganguli, S. S., Abuseda, H., & Debnath, A. (2021). Petrophysical assessment and permeability modeling utilizing core data and machine learning. *Journal of Petroleum Science & Engineering*.
- Foote, K. D. (2019, March 26). A Brief History of Machine Learning. *Dataversity*.
- Ford, M. (2018). *Architects of Intelligence: The Truth About AI from the People Building It*. Packt Publishing.
- Gamal, H., & Elkatatny, S. (2022). Prediction model based on an artificial neural network for rock porosity. *Arabian Journal for Science and Engineering*, *47*(9), 11211–11221.
- Halliburton. (2020). *The challenges of well log-based petrophysical estimation: Variability, calibration, and uncertainty*. Halliburton Technical Report.
- Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
- Hao, Z., Ying, C., Su, H., Zhu, J., Song, J., & Cheng, Z. (2021). Physics-informed neural networks.
- Hassaan, S., Mohamed, A., Ibrahim, A. F., & Elkatatny, S. (2024). Real-Time Prediction of Petrophysical Properties Using Machine Learning Based on Drilling Parameters. *ACS Omega*, *9*(15), 17066–17075.
- Helle, H. B., Bhatt, A., & Ursin, B. (2001). Porosity and permeability prediction from wireline logs using artificial neural networks: A North Sea case study. *Geophysical Prospecting*, *49*(4), 431–444.
- Hyne, N. J. (2012). *Nontechnical Guide to Petroleum Geology, Exploration, Drilling, and Production* (3rd ed.). PennWell.

Bibliography

- Johnson, R., & Wang, L. (2020). Machine learning approaches for porosity prediction in petroleum reservoirs: A review of conventional logging techniques. *Energy Reports*, *6*(5), 302–315.
- Kabir, M. A., & Luo, X. (2020). Unsupervised learning for network flow-based anomaly detection in the era of deep learning. In *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 165–168). IEEE.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.
- Kamali, M. Z., Davoodi, S., Ghorbani, H., Wood, D. A., Mohamadian, N., Lajmorak, S., ... & Band, S. S. (2022). Permeability prediction of heterogeneous carbonate gas reservoirs.
- Kermia, H., Ameur-Zaimeche, O., Belksier, M. S., Boutaghane, A., Tahar-Belkacem, N., Ouladmansour, A., & Kechiched, R. (2025). Advanced machine learning and mud logging data for dynamic Poisson's ratio prediction while drilling in tight fractured reservoir of Quartzite El Hamra, Rhourde Nouss field, Algeria. *Petroleum Science and Technology*, 1–19. https://doi.org/10.1080/10916466.2025.2477648
- Khassaf, Ahmed K., Al-hameed, Zainab M., Al-Mohammedawi, Noor R., Al-Mudhafar, Watheq J., Wood, David A., Abbas, Mohammed A., Ameur-Zaimeche, Ouafi, and Ahmed A. Alsubaih. "Physics-Informed Machine Learning for Enhanced Permeability Prediction in Heterogeneous Carbonate Reservoirs." Paper presented at the Offshore Technology Conference, Houston, Texas, USA, May 2025. doi: https://doi.org/10.4043/35892-MS
- Lee, K.-F. (2018). *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Harcourt.
- Lucia, F. J. (2003). Petrophysical Parameters and Their Role in Reservoir Characterization. In *Carbonate Reservoir Characterization* (pp. 47–68). Springer.
- MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
- Makhous, M., Aït-Laoussine, N., & Mebarki, N. (2009). Hydrocarbon systems of the Algerian Sahara: Structural evolution and petroleum potential. *Journal of Petroleum Geology*, *32*(2), 123–144.
- Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. III. (1972). *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Universe Books.
- Meadows, D. H., Randers, J., & Meadows, D. L. (2004). *The Limits to Growth: The 30-Year Update*. Chelsea Green Publishing Co.
- Mellal, I., Malki, M. L., Latrach, A., Ameur-Zaimeche, O., & Bakelli, O. (2023, June). Multiscale Formation Evaluation and Rock Types Identification in the Middle Bakken Formation. In *SPWLA Annual Logging Symposium* (p. D031S002R006). SPWLA. https://doi.org/10.30632/SPWLA-2023-0012
- Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Mohammadian, E., Kheirollahi, M., Liu, B., Ostadhassan, M., & Sabet, M. (2022). A case study of petrophysical rock typing and permeability prediction using machine learning in a heterogenous carbonate reservoir in Iran. *Scientific Reports*, *12*, 4505.

Bibliography

- Mohammed, M., Khan, M. B., & Bashier, M. B. E. (2016). *Machine Learning: Algorithms and Applications*. CRC Press.
- Nilsson, N. J. (2009). *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge University Press.
- Ouladmansour, A., Ameur-Zaimeche, O., Kechiched, R., Heddam, S., & Wood, D. A. (2023). Integrating drilling parameters and machine learning tools to improve real-time petrophysical predictions.
- Ouladmansour, A., Ameur-Zaimeche, O., Kechiched, R., Heddam, S., & Wood, D. A. (2023). Integrating drilling parameters and machine learning tools to improve real-time porosity prediction of multi-zone reservoirs. Case study: Rhourd Chegga oilfield, Algeria. *Geoenergy Science and Engineering*, *223*, 211511. https://doi.org/10.1016/j.geoen.2023.211511
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
- Qalandari, R., Zhong, R., Salehi, C., Chand, N., Johnson, R. L., Vazquez, G., ... & Zimmerman, J. (2022, October). Estimation of Rock Permeability Scores Using Machine Learning Methods. In *SPE Asia Pacific Oil and Gas Conference and Exhibition* (p. D021S008R001). SPE.
- Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*, 59–66.
- Saba, C. S., & Ngepah, N. (2022). Convergence in renewable energy sources and the dynamics of their determinants: An insight from a club clustering algorithm. *Energy Reports*, *8*, 3483–3506.
- Sahraoui, S., & Mahdjoudi, O. R. (2024). *Étude sédimentologique et pétrophysique au niveau du réservoir Ordovicien dans le bassin de Oued Mya (champ Hassi Tarfa)*. Master's Thesis, Université Kasdi Merbah, Ouargla.
- Saiad, M., & Derbal, O. (2016). *Caractérisation pétrographique et structurale des grès ordoviciens de Quartzite El Hamra (champ de Hassi Tarfa)*. Master's Thesis, Université Kasdi Merbah, Ouargla.
- Santosh, K., Trivedi, M., Al-Jarrah, O., & Aloysius, N. (2020). Artificial intelligence in oil and gas: A review of applications and challenges. *Journal of Petroleum Science and Engineering*, *191*, 107209.
- Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from a machine learning perspective. *Journal of Big Data*, *7*(1), 1–29.
- Schlumberger. (2019). *Formation heterogeneity and its impact on well log reliability: A case study of Middle Eastern carbonate reservoirs*. Schlumberger Reservoir Review.
- Seber, G. A., & Lee, A. J. (2003). *Linear Regression Analysis* (Vol. 330). John Wiley & Sons.
- Sharma, A., Burak, T., Nygaard, R., Hoel, E., Kristiansen, T., Hellvik, S., & Welmer, M. (2023, September). Projecting Petrophysical Logs at the Bit through Multi-Well Data Analysis with Machine Learning. In *SPE Offshore Europe Conference and Exhibition* (p. D031S012R001). SPE.

Bibliography

- Sharma, P., Gupta, R., & Al-Ghamdi, M. (2023). Hybrid machine learning models for permeability prediction: Integrating well logs and real-time drilling data. *Journal of Natural Gas Science and Engineering*, *114*, 104021.
- Smith, J., & Roberts, B. (2019). Machine learning techniques for permeability estimation: A comparison of well log-based methods. *Petroleum Geoscience*, *25*(3), 244–258.
- Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294.
- Tian, J., Qi, C., Sun, Y., Yaseen, Z. M., & Pham, B. T. (2021). Permeability prediction of porous media using a combination of computational fluid dynamics and hybrid machine learning methods. *Engineering with Computers*, *37*, 3455–3471.
- Tiab, D., & Donaldson, E. C. (2015). *Petrophysics: Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties*. Gulf Professional Publishing.
- Wang, J., Yan, W., Wan, Z., Wang, Y., Lv, J., & Zhou, A. (2020). Prediction of permeability using random forest and genetic algorithm model. *Computer Modeling in Engineering & Sciences*, *125*(3), 1135–1157.
- Zolotukhin, A. B., & Gayubov, A. T. (2019, November). Machine learning in reservoir permeability prediction and modelling of fluid flow in porous media. In *IOP Conference Series: Materials Science and Engineering* (Vol. 700, No. 1, p. 012023). IOP Publishing.
- Zhang, Y., Chen, X., & Liu, P. (2023). Advances in machine learning applications for petrophysical analysis: A systematic review. *Computers & Geosciences*, *165*, 104879.