Kasdi Merbah University – Ouargla

Faculty of Hydrocarbons, Renewable Energies

and Earth and Universe Sciences

Department of Earth and Universe Sciences

Professional Master's Thesis

Domain: Earth and Universe Sciences

Sector: Geology

Specialty: Petroleum Geology

Theme

Artificial Intelligence-Based Prediction of Flow Zone Indicator Using Mud Logging Data

Presented by:

- **Taleb Mohamed Brahim**
- **Belmiloud Toufik**
- **Sendid Yahia**

Publicly defended on: 00/06/2025 **In front of the jury:**

President:

Examiner:

Examiner:

Supervisor: AMEUR-ZAIMECHE Ouafi MCA Univ. Ouargla

Academic year: 2025/2026

ACKNOWLEDGEMENT

First and foremost, we express our deepest gratitude to Allah Almighty for granting us the strength, patience, and perseverance to complete this work successfully.

We would like to extend our sincere appreciation to our supervisor, ***Dr. Ameur-Zaimeche Ouafi and Bouthaghane Ayoub****, for his invaluable guidance, continuous support, insightful advice, and encouragement throughout the realization of this dissertation. His expertise and dedication have greatly contributed to the completion of this research.*

Our heartfelt thanks also go to the esteemed members of the jury, …….for accepting to evaluate this work and for their valuable comments and recommendations.

We are grateful to all the professors and staff members of the Department of Earth and Universe Sciences for the knowledge, assistance, and support they provided during our academic journey.

We would also like to acknowledge our colleagues and friends whose encouragement, cooperation, and friendship have been a source of motivation throughout this project.

Finally, we extend our sincere gratitude to our families for their unconditional love, patience, understanding, and constant support. Their encouragement has been an essential source of strength throughout our studies.

To everyone who contributed, directly or indirectly, to the successful completion of this work, we offer our sincere thanks and appreciation.

DEDICATION

I dedicate this work with deep love and gratitude:

To my beloved parents, *whose endless love, sacrifices, encouragement, and prayers have always been my greatest source of strength. This achievement would not have been possible without their unwavering support and guidance.*

To my brothers and sisters*, for their affection, understanding, and constant encouragement throughout my academic journey.*

To my family*, whose confidence in me has inspired me to pursue my goals with determination and perseverance.*

To my friends and colleagues*, who shared memorable moments, offered support, and contributed to making this journey both rewarding and enjoyable.*

To all my teachers and professors*, who have enriched my knowledge and helped shape my academic and personal development.*

Finally, to everyone who has supported me, directly or indirectly, in the completion of this work.

I dedicate this dissertation to you all.

Taleb Mohamed Brahim

DEDICATION

I dedicate the fruits of this labor, first and foremost, to the inexhaustible source of giving ***my dear parents*** to those who paved my path to success with them unconditional love, and who were my safe haven and my greatest support. Your immense sacrifices and unwavering faith in me are the cornerstone of every ambition I have achieved, and my constant source of inspiration.

To my support and strength, ***my brothers and sisters,*** who surrounded me with patience and encouragement during the most challenging moments of this journey.

To my companions on this path, ***my friends and colleagues,*** with whom I shared the passion for learning and the burden of challenges, and who were my best companions, giving this journey value and meaning.

To the shapers of minds, ***my esteemed teachers and professors,*** who spared no effort or guidance, and who were the guiding light that illuminated my path and refined my knowledge.

And finally, to every passionate and persevering individual who understands that unwavering will and diligent work are the gateways to achievement.

Thank you all for being a part of this journey.

Belmiloud Toufik

DEDICATION

To my beloved parents, whose love, prayers, and unwavering sacrifices were the foundation of my success.

To my brothers and sister, and to all my friends who have accompanied me throughout my educational journey, and to everyone who has taught me even a single letter,

Thank you for sharing the challenges, the growth, and the unforgettable moments.

Sendid Yahia

Summary

CHAPTER 01:LITERATURE REVIEW ON ARTIFICIAL INTELLIGENCE METHODS

Chapter 02: Geological and Reservoir Overview of the Study Area

Chapter 03 : Materials and Methods

Chapter 04 : Results & Discussions

Abstract

Reservoir characterization is a crucial step in evaluating petroleum reservoirs, especially in heterogeneous formations that are difficult to accurately represent using limited core data. This work aims to predict the real-time Flow Zone Indicator (FZI) using mud logging data, specifically drilling data (DD). The study employed supervised machine learning models, Random Forest, Gradient Boosting, Gradient Boosting Regressor, SVM, and, to predict FZI values. SHAP analysis was also used to identify the most significant variables affecting the model predictions.

Using key statistical metrics including the coefficient of determination (R^2) and root mean square error (RMSE), the results demonstrated that the Random Forest and Gradient Boosting gave the best predictive performance, while SVM showed weaker results. The best performance was obtained with the reduced input combination in RF Scenario 4 with an R^2 of 0.809884, and RMSE of 3.789239.

SHAP analysis identified ROP, RPM, Q, and Torque as the most influential variables for FZI prediction. These findings confirm the effectiveness of integrating real-time drilling data with machine learning algorithms to provide a reliable and rapid alternative for estimating reservoir quality, thereby reducing reliance on costly laboratory data and supporting better operational decision-making during drilling.

Keywords:

Flow Zone Indicator (FZI)
Machine Learning
Mud Logging Data
Random Forest
Reservoir Characterization
Explainable AI (SHAP)
TAGI Reservoir

Résumé

La caractérisation des réservoirs est une étape cruciale de l'évaluation des gisements pétroliers, notamment dans les formations hétérogènes difficiles à représenter avec précision à partir de données de carottes limitées. Ce travail vise à prédire en temps réel l'indice de zone d'écoulement (FZI) à partir de données de diagraphie de boue, en particulier les données de forage (DD) L'étude a utilisé des modèles d'apprentissage

automatique supervisé, Random Forest, Gradient Boosting , SVM , pour prédire les valeurs du FZI. La technologie SHAP a également été employée pour identifier les variables les plus significatives influençant les prédictions du modèle.

En utilisant des métriques statistiques clés, notamment le coefficient de détermination (R^2) et l'erreur quadratique moyenne (RMSE), Les résultats ont démontré que les algorithmes Random Forest et Gradient Boosting ont fourni les meilleures performances prédictives, tandis que le modèle SVM (Support Vector Machine) a présenté des résultats moins satisfaisants. La meilleure performance a été obtenue avec la combinaison réduite de variables d'entrée dans le Scénario 4 du modèle Random Forest, atteignant un coefficient de détermination (R^2) de 0,809884 et une erreur quadratique moyenne (RMSE) de 3,789239.

SHAP a permis d'identifier les variables les plus influentes dans la prédiction des valeurs de FZI, qui sont (ROP, RPM, Q et Torque). Ces résultats confirment l'efficacité de l'intégration des données de forage en temps réel avec les algorithmes d'apprentissage automatique pour fournir une alternative fiable et rapide à l'estimation de la qualité du réservoir, réduisant ainsi la dépendance aux données de laboratoire coûteuses et favorisant une meilleure prise de décision opérationnelle lors du forage.

Mots-clés :

Indicateur des zones d'écoulement (FZI)
Apprentissage automatique
Données de Mud Logging
Forêt aléatoire (Random Forest)
Caractérisation des réservoirs
Intelligence artificielle explicable (SHAP)
Réservoir TAGI (Trias Argilo-Gréseux Inférieur)

ملخص

تُعدّ عملية توصيف المكامن من أهم المراحل في تقييم الخزانات البترولية، خاصةً في التكوينات الغير المتجانسة التي يصعب تمثيلها بدقة اعتمادًا على بيانات اللب المحدودة. يهدف هذا العمل إلى التنبؤ بمؤشر منطقة التدفق FZI في الزمن الحقيقي باستخدام بيانات تسجيل الطين، المتمثلة في بيانات الحفرDD، اعتمدت الدراسة على تطبيق نماذج التعلم الآلي الخاضع للإشراف، Random Forest و Gradient Boosting و SVM وANN ، للتنبؤ بقيم FZI, كما استعملت تقنية SHAP لتحديد أهم المتغيرات المؤثرة في تنبؤات النموذج

باستخدام مقاييس إحصائية رئيسية شملت معامل التحديد (R^2)، ومتوسط مربع الخطأ الجذري (RMSE) أظهرت النتائج أن خوارزميتي الغابة العشوائية (Random Forest) والتعزيز التدريجي (Gradient Boosting) حققتا أفضل أداء تنبؤي، في حين أظهر نموذج آلة المتجهات الداعمة (SVM) نتائج أقل كفاءة. وقد تم تحقيق أفضل أداء باستخدام المجموعة المختزلة من متغيرات الإدخال في السيناريو الرابع لنموذج الغابة العشوائية، حيث بلغ معامل التحديد0.809884 (R^2) ، بينما بلغت قيمة الجذر التربيعي لمتوسط مربع الخطأ.3.789239 (RMSE)

ومن خلال تحليل SHAP، تم تحديد أن المتغيرات الأكثر تأثيراً في التنبؤ بقيم FZI هي (ROP, RPM, Torque, Q) تؤكد هذه النتائج فعالية دمج بيانات الحفر في الوقت الفعلي مع خوارزميات التعلم الآلي لتوفير بديل موثوق وسريع لتقدير جودة المكمن، مما يقلل من الاعتماد على البيانات المختبرية المكلفة ويدعم اتخاذ قرارات تشغيلية أفضل أثناء الحفر.

الكلمات المفتاحية:
مؤشر مناطق الجريان(FZI)
التعلم الآلي
بيانات تسجيل الطين(Mud Logging Data)
الغابة العشوائية(Random Forest)
توصيف الخزانات البترولية
الذكاء الاصطناعي القابل للتفسير(SHAP)
خزان التریاس السفلي الطيني الرملي(TAGI)

List of figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANFIS	Adaptive Neuro-Fuzzy Inference System
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CS	Computer Science
CNNs	Convolutional Neural Networks
DD	Drilling Data
DL	Deep Learning
DNN	Deep Neural Networks
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Tree
FZI	Flow Zone Indicator
GB	Gradient Boosting
GWD	Gas While Drilling
GLT	Geochemical Logging Tools
HFUs	Hydraulic Flow Units
ML	Machine Learning
MAE	Mean Absolute Error
MDT	Modular Formation Dynamics Tester
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
RQI	Reservoir Quality Index
RF	Random Forest

RL	Reinforcement Learning
RFT	Repeat Formation Tester
RCAL	Routine Core Analysis
R	Correlation Coefficient
R^2	coefficient of determination
RMSE	Root Mean Square Error
RNNs	Recurrent Neural Networks
SVM	Support Vector Machine
SML	supervised machine learning
SAG	Silurian Clay-Sand Reservoirs
SF	Sif Fatima
TAC	Clay-carbonate Triassic
TAGI	Lower Clay-sand Triassic
TAGS	Upper Clay-sand Triassic
VCL	Volume of Clay
XGBoost	Extreme Gradient Boosting

GENERAL INTRODUCTION

GENERAL INTRODUCTION

The oil and gas industry remains a strategic sector for global energy supply, where improving reservoir evaluation and production efficiency is essential for reducing uncertainty and optimizing field development. In this context, reservoir characterization represents a fundamental step because it helps describe the spatial distribution of petrophysical properties and supports decisions related to drilling, completion, production planning, and reservoir management. However, many reservoirs are highly heterogeneous, meaning that their porosity, permeability, pore-throat distribution, and fluid-flow capacity can vary significantly from one interval to another. (Tiab et al., 2022).

The conventional calculation of FZI is generally based on core-derived porosity and permeability. First, the Reservoir Quality Index (RQI) is calculated from permeability and effective porosity, then FZI is obtained by dividing RQI by the normalized porosity index. This approach is physically meaningful because it is derived from the modified Kozeny–Carman concept, which relates permeability to pore geometry and porosity. Therefore, FZI reflects the influence of pore-throat size, pore connectivity, tortuosity, and depositional or diagenetic processes on fluid flow inside the reservoir. (Kozeny, 1927; Carman, 1997; Amaefule et al., 1993; Sadeghpour et al., 2025).

Despite its usefulness, the conventional FZI method has important limitations. Core analysis provides accurate measurements, but core samples are expensive, time-consuming, discontinuous, and limited to selected intervals. As a result, the direct use of core data alone is often insufficient for continuous reservoir characterization, especially in uncored wells or intervals. (Zhu et al., 2018)

To overcome these limitations, recent studies have focused on the use of indirect and data-driven methods to estimate reservoir quality parameters. Well logs have been widely used because they provide continuous information along the wellbore, and several studies have shown that petrophysical logs can be used to estimate FZI with good accuracy. (Mehrabi et al., 2020)

In recent years, Artificial Intelligence and Machine Learning have become powerful tools in petroleum engineering because they can model complex nonlinear relationships between input and output variables. Machine learning algorithms such as Random Forest, Gradient Boosting, XGBoost, Support Vector Machine, and Artificial Neural Networks have

been successfully applied to predict petrophysical properties, lithology, well logs, permeability, porosity, and reservoir quality indicators. These methods are especially useful when conventional empirical relationships are not sufficient to describe complex heterogeneous formations. (Lawal et al., 2024)

However, one of the main challenges of machine learning models is their interpretability. In petroleum engineering applications, it is not enough to obtain accurate predictions; it is also necessary to understand why the model gives a certain result and which input variables control the prediction. For this reason, explainable machine learning methods, especially SHAP analysis, are increasingly used to interpret model behavior. SHAP helps quantify the contribution of each input parameter to the predicted FZI value and allows engineers to identify the most influential drilling or mud logging variables. (Sadeghpour et al., 2025).

The objective of this study is to develop an explainable machine learning framework for predicting the Flow Zone Indicator using mud logging data. The proposed approach aims to use real-time drilling and gas data as input variables to estimate FZI without relying only on core measurements.

Predicting FZI from mud logging data can help identify productive intervals, classify hydraulic flow units, estimate reservoir quality in uncored zones, and support faster operational decisions during drilling. In addition, the integration of explainable machine learning improves confidence in the results by transforming the predictive model from a black-box tool into an interpretable decision-support system. (Amaefule et al., 1993).

Overall, the prediction of FZI using mud logging data and explainable machine learning represents a modern and efficient approach for reservoir characterization. It reduces dependence on costly and limited core data, improves the use of real-time drilling information, and provides a better understanding of reservoir heterogeneity. This approach can support better reservoir management, optimize field development, and contribute to safer and more cost-effective hydrocarbon production. (Sadeghpour et al., 2025).

Thesis Organization

This thesis is organized into four main chapters, designed to cover the theoretical, geological, methodological, and practical aspects of Flow Zone Indicator prediction using machine learning techniques.

Chapter One: Literature Review on Artificial Intelligence Methods

This chapter presents the basic concepts of artificial intelligence, machine learning, and deep learning. It explains supervised and unsupervised learning methods, artificial neural networks, and the main applications of artificial intelligence in petroleum engineering, especially in exploration, seismic interpretation, drilling, and reservoir characterization.

Chapter Two: Geological and Reservoir Overview of the Study Area

This chapter provides a geological description of the Berkine Basin and the Sif Fatima Field. It presents the geographical location, stratigraphic framework, petroleum system, structural setting, and reservoir characteristics, with particular focus on the TAGI reservoir and its fluvial facies distribution.

Chapter Three: Materials and Methods

This chapter describes the conventional and unconventional methods used in reservoir characterization. It explains porosity, permeability, Hydraulic Flow Units, and Flow Zone Indicator concepts. It also presents the machine learning workflow, including data collection, data cleaning, descriptive statistics, data splitting, algorithm selection, and validation criteria.

Chapter Four: Results and Discussions

This chapter presents the results obtained from statistical analysis and machine learning models. It includes univariate and bivariate analysis, model performance comparison, cross-plots, feature importance, and SHAP interpretation. The chapter discusses the ability of drilling data to predict FZI and identifies the best-performing model and input combination.

Chapter 01
Literature Review on Artificial Intelligence Methods

Introduction

Artificial Intelligence (AI) has emerged as a transformative technology in recent decades, fundamentally changing the way complex systems are analyzed and managed. It is generally defined as the capability of machines to perform tasks that typically require human intelligence, including learning, reasoning, and decision-making. Core components of AI, such as machine learning (ML) and deep learning (DL), rely on data-driven models and artificial neural networks to identify patterns and extract meaningful information from large datasets (Russell & Norvig, 2021; Goodfellow et al., 2016).

These approaches have significantly improved the ability to handle high-dimensional and non-linear problems across various scientific and engineering disciplines. AI has found widespread applications in numerous fields, including healthcare, finance, transportation, and industrial systems. In healthcare, AI is utilized for diagnostic imaging and disease prediction, while in finance it supports risk assessment and automated trading systems. In transportation, it enables the development of autonomous vehicles, and in industrial environments, it contributes to process optimization and predictive maintenance. The adaptability of AI techniques allows them to be applied across diverse domains, enhancing operational efficiency and decision-making capabilities (Jordan & Mitchell, 2015; LeCun et al., 2015).

In recent years, the oil and gas industry has increasingly adopted AI technologies to address the growing complexity of exploration and production processes. The sector generates vast amounts of data from seismic surveys, well logging, drilling operations, and production monitoring. Traditional analytical methods are often insufficient to fully exploit these data. AI-based approaches provide advanced tools for data integration, interpretation, and prediction, enabling more accurate and faster decision-making in reservoir and field management (Mohaghegh, 2018; Aminzadeh & de Groot, 2006).

Within petroleum engineering, AI plays a crucial role in reservoir characterization, where it is used to estimate key petrophysical properties such as porosity, permeability, and fluid saturation. Machine learning models improve reservoir simulation and forecasting by capturing complex relationships between geological and dynamic variables. In drilling engineering, AI techniques are applied to optimize well trajectories, detect anomalies, and reduce operational risks. Furthermore, in production engineering, AI supports real-time monitoring, predictive maintenance of equipment, and production

optimization, ultimately enhancing recovery efficiency and operational safety (Al-Bulushi et al., 2017; Holzmann et al., 2020).

Consequently, the integration of AI into the oil and gas industry represents a significant advancement, offering new opportunities to improve efficiency, reduce costs, and maximize hydrocarbon recovery. As the industry faces increasing technical, economic, and environmental challenges, AI is expected to play a central role in driving innovation and supporting sustainable energy development (Ahmed et al., 2019; Mohaghegh, 2020).

1 Concepts of Artificial Intelligence and Machine Learning

1.1 Artificial Intelligence

Artificial intelligence (AI) is the ability of machines to replicate or enhance human intellect, such as reasoning and learning from experience. AI's ability to unearth valuable information among layers of data demonstrates its benefits. Artificial Intelligence (AI) is the fascinating field of combining human intelligence and computing power to generate smart or intelligent solutions to complex problems. AI-based solutions are capable of not only simulating human brain functionality but are also adept at using intelligent algorithms to understand the problem and provide solutions. Ever since the early 1980s, AI has been integrated into various business applications. the Literature shows that AI has been spreading in all industrial applications and providing fast, reliable and easy-to-use solutions to complex problems in this time of Fourth Industrial Revolution (IR4.0).

Interestingly, the areas of research in AI have been growing many folds, with techniques penetrating new avenues and industrial applications (King Fahd University of Petroleum & Minerals, 2021).

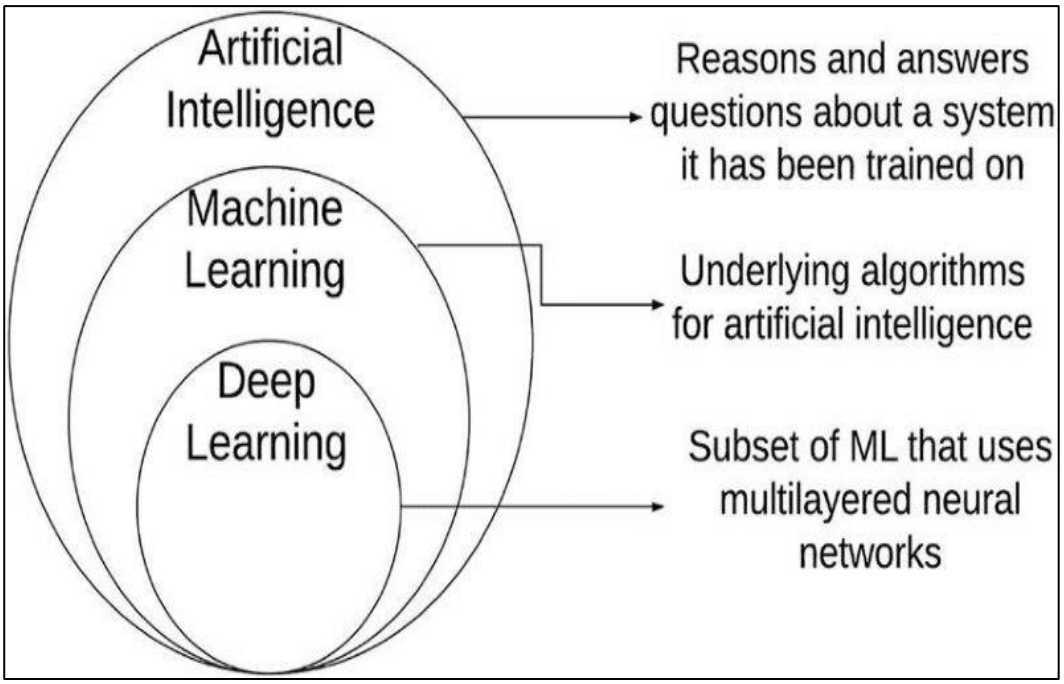


Figure 1. Venn diagram showing relationship between Artificial Intelligence, Machine Learning and Deep Learning (Bengio; Goodfellow; Courville, 2015).

1.2 Machine Learning

Machine Learning (ML) originates from computer science (CS) and Artificial Intelligence (AI), which address systems that can learn from data, rather than merely following the programmed instructions explicitly. Moreover, ML has a close relationship with optimization and statistics, delivering both theory and methods to the field. ML is applied to various computing tasks in which designing and programming rule-based, explicit algorithms is impractical. In some situations, ML, pattern recognition, and data mining are conflated. Arthur Samuel, in 1959, defined ML as a field of study that gives computers the ability to learn without being clearly programmed. In general, ML and data mining are confused with each other since they use the same methods and they significantly overlap. These two can be described as follows. ML is focused on prediction based on recognized properties that are learned from training data. On the other hand, data mining is concentrated on discovering (previously) unknown properties in the data. The two areas are overlapped in many aspects: in data mining, numerous machine learning methods are used, but with a diverse goal in mind. However, ML makes use of data mining methods as “unsupervised learning” or as a preprocessing step for the .improvement of learner accuracy. The present research is focused on ML (Mitchell, 1997)

1.2.1 Types of Machine Learning

Machine learning algorithms come in a variety of forms, each with unique advantages and disadvantages (Alpaydin, 2010). Some of the machine learning algorithms that are most frequently employed are listed below:

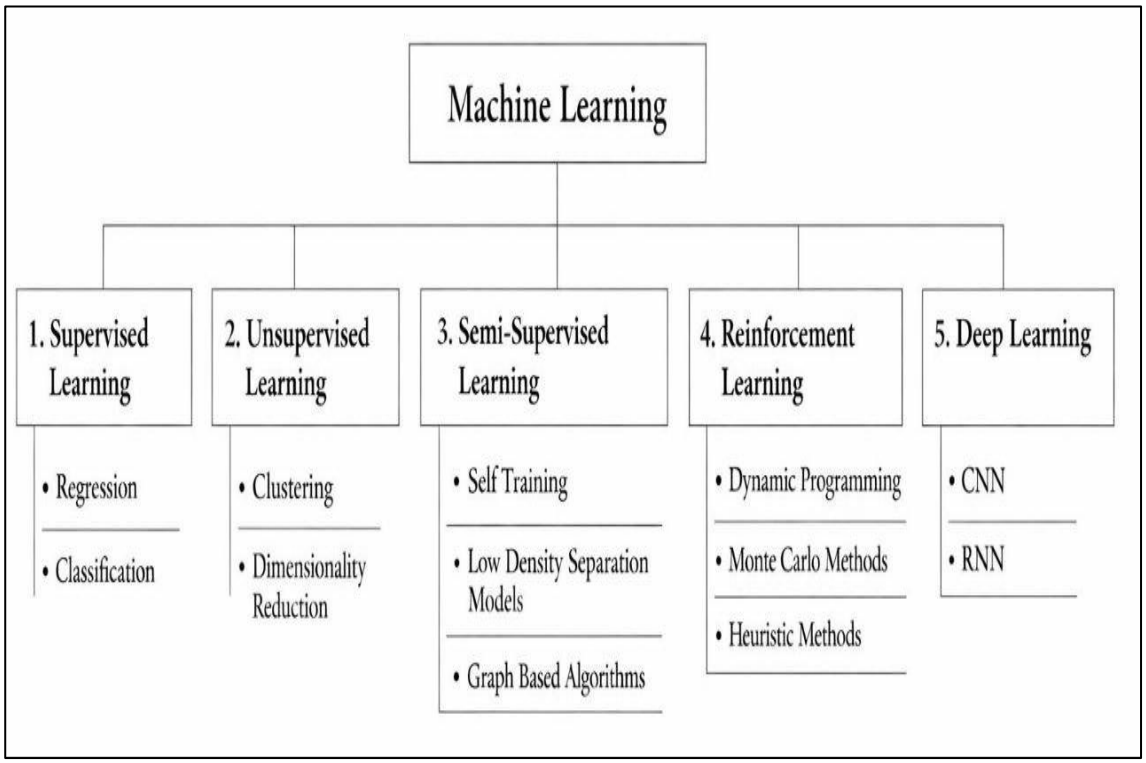


Figure 2. Diagram showing the different types of machine learning (from Nassif et al., 2019) modified.

1.2.2 History of Machine Learning Algorithms in Petroleum Geology

Petroleum geology is one of many domains to which machine learning (ML) algorithms have been applied for many years. The history of ML algorithms in petroleum geology is briefly described:

1980s–1990s: Researchers used methods like artificial neural networks (ANNs) to categorize lithofacies and estimate porosity and permeability in reservoirs during this time, which is when ML algorithms were first applied to petroleum geology (Rogers et al., 1989; Xie et al., 1994; Pwavodi et al., 2023).

Support vector machines (SVMs) and decision trees were used in the 2000s to analyze petrophysical and seismic data in order to predict reservoir characteristics, lithofacies, and hydrocarbon indications (Chen et al., 2001; Wu et al., 2008; Oguadinma

et al., 2016; Oguadinma et al., 2017; Nwaezeapu et al., 2018; Oguadinma et al., 2021; Ibekwe et al., 2023).

2010s: has seen a rise in the use of machine learning (ML) techniques in petroleum geology due to the development of big data and cloud computing. Seismic data has been subjected to deep learning methods for fault detection and classification, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Liu et al., 2016; Liu et al., 2018). For lithofacies classification and reservoir property prediction, ensemble learning approaches like random forests and gradient boosting have been applied (Wang et al., 2017; Xu et al., 2018).

At this time, ML algorithms are still being developed and used for a variety of petroleum geology-related tasks, such as reservoir characterization, drilling optimization, and well log interpretation. Also, researchers are investigating the application of ML algorithms for production optimization and preventative maintenance. (Abdelaziz et al., 2021; Wang et al., 2021; Ibekwe et al., 2023)

1.2.3 Supervised Learning

The learning process in a simple machine learning model is divided into two steps: training and testing. In training process, samples in training data are taken as input in which features are learned by learning algorithm or learner and build the learning model. In the testing process, learning model uses the execution engine to make the prediction for the test or production data. Tagged data is the output of learning model which gives the final prediction or classified data. Supervised learning is the most common technique in the classification problems, since the goal is often to get the machine to learn a classification system that we've created. Most commonly, supervised learning leaves the probability of input undefined, such as an input where the expected output is known. In the training process provides datasets consisting of features and labels. The main task is to construct an estimator able to predict the label of an object given by the set of features. Then, the learning algorithm receives a set of features as inputs along with the correct outputs and it learns by comparing its actual output with corrected outputs to find errors. It then modifies the model accordingly. The model that is created is not needed if the inputs are an overview of the supervised machine learning methods available, but if some of the input values are missing, it is not possible to infer anything about the outputs. As

previously mentioned, the supervised learning tasks are divided into two categories: classification and regression. In classification, the label is discrete, while in regression, the label is continuous. (Nasteski, 2017).

a) Classification

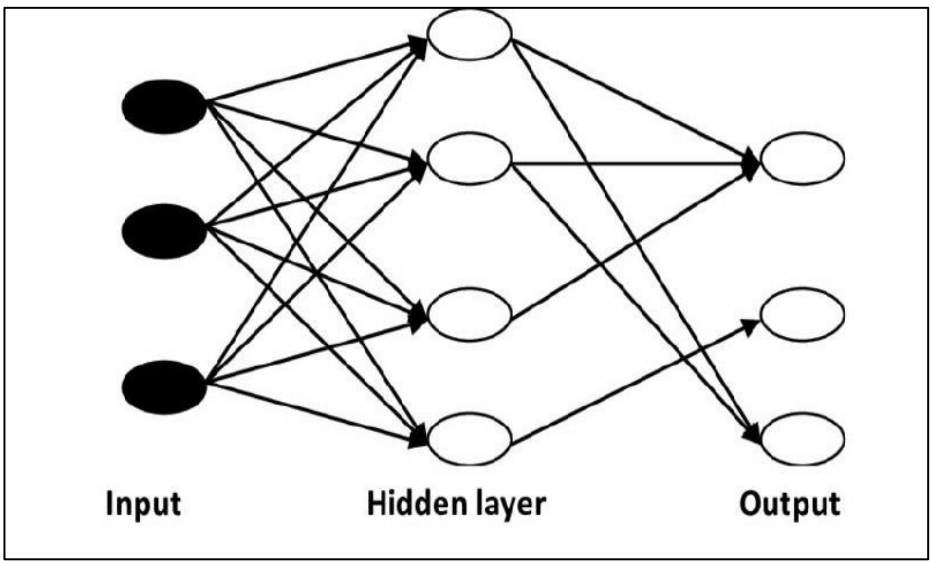


Figure 3. A Simple neural network (Anirbid et al 2021) modified

The algorithm makes the distinction between the observed data X that is the training data, in most cases structured data given to the model during the training process. In this process, the supervised learning algorithm builds the predictive model. After its training, the fitted model would try to predict the most likely labels for a new set of samples X in the testing set. Depending on the nature of the target y, supervised learning can be classified If y has values in a fixed set of categorical outcomes (integers), the task to predict y is called classification.

b) Regression

The algorithm is making the same process, but if y has floating point values, the task to predict y is called regression.

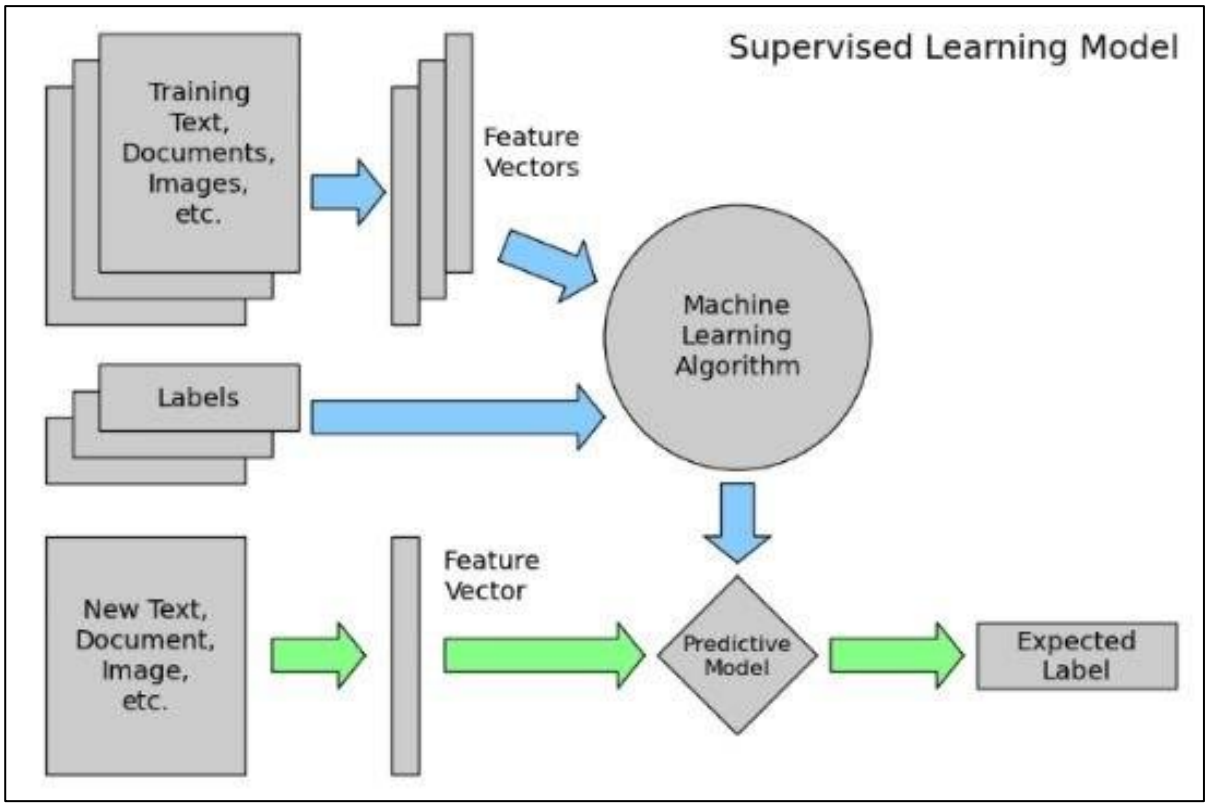


Figure 4. supervised learning model (https://www.researchgate.net)

1.2.3.1 Supervised Learning algorithmes

When choosing an SML algorithm, the heterogeneity, precision, excess, and linearity of the information ought to be examined before selecting an algorithm. SML is used in a various range of applications such as speech and object recognition, bioinformatics, and spam detection. Recently, advances in SML are being witnessed in solid-state material science for calculating material properties and predicting their structure. This review covers various algorithms and real-world applications of SML. The key advantage of SML is that, once an algorithm swots with data, it can do its task automatically. (Shetty., et al. 2022).

1.2.3.1.1 Decision trees

Decision tree represents a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form so called root tree, which means that it is a distributed tree with a basic node called root with no incoming edges. All the other nodes have exactly one incoming edge. The node that has outgoing edges is called internal node or a test node. The rest of the nodes are called leaves. In a decision tree, each test node splits the instance space into two or more sub-spaces according to a

certain discrete function of the input values. In the simplest case, each test considers a single attribute, such that the instance space is portioned according to the attribute's value. In case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class that represents the most appropriate target value. The leaf may hold a probability vector that indicates the probability of the target attribute having a certain value. The instances are classified by navigating them from the root of the tree down the leaf, according to the outcome of the tests along the path. On Figure describes a simple use of the decision tree. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values. Given this classifier, the analyst can predict the response of some potential customers and understand the behavioral characteristics of the entire potential customers population. (Nasteski, 2017).

1.2.3.1.2 Linear regression

The goal of linear regression, as a part of the family of regression algorithms, is to find relationships and dependencies between variables. It represents a modeling relationship between a continuous scalar dependent variable y (also referred to as a label or target in machine learning terminology) and one or more (a D-dimensional vector) explanatory variables (also called independent variables, input variables, features, observed data, observations, attributes, dimensions, data points, etc.), denoted as X, using a linear function. In regression analysis, the goal is to predict a continuous target variable, whereas another area called classification is concerned with predicting a label from a finite set. (Nasteski, 2017).

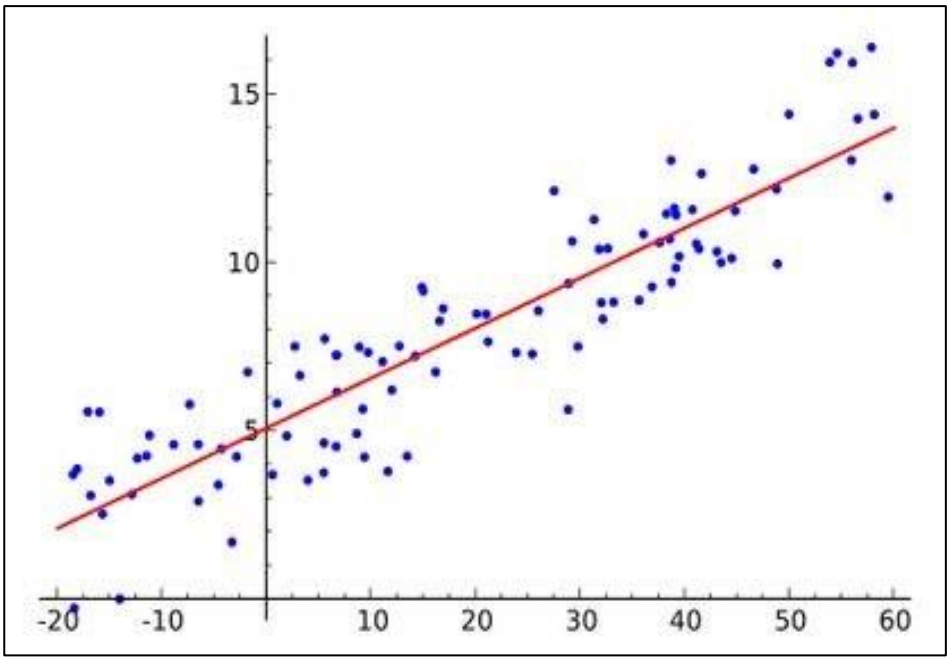


Figure 5. Visual representation of the linear regression (https://cdn.activestate.com)

As shown in Figure, the model (red line) is calculated using training data (blue points), where each point has a known label (y-axis), to fit the points as accurately as possible by minimizing the value of a chosen loss function. We can then use the model to predict unknown labels (where we only know the x value and want to predict the y value).

1.2.3.1.3 Random Forest (RF)

Random Forests is a machine learning technique that employs an ensemble of decision trees to improve classification accuracy and reduce prediction errors. It works by combining the outputs of multiple individual trees, which together produce more robust and stable results than a single decision tree. (Breiman,2001).

$$\widehat{Y} = \boldsymbol{\phi}_{T \cdot p}(X) = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\phi} T_{b \cdot m}(X)$$

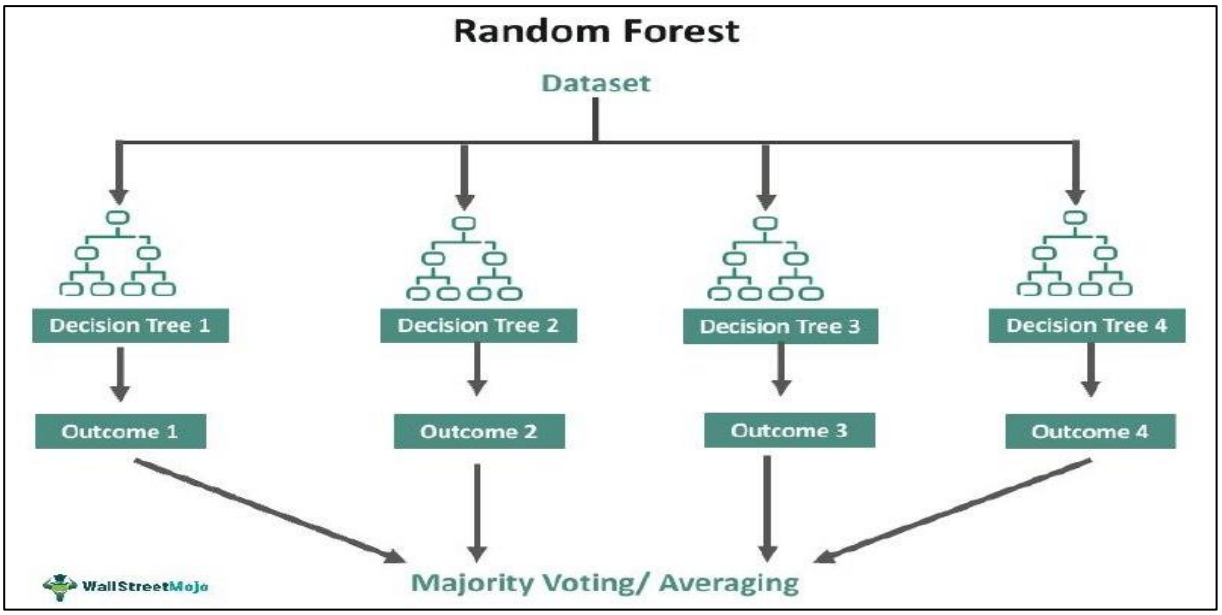


Figure 6. Random Forest model (https://wallstreetmojocms.recurpro.in/uploads)

1.2.3.1.4 XGboost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm based on tree boosting. It is designed to be scalable and highly efficient for processing large datasets and achieving accurate results. It is widely used by data scientists in competitions and real-world applications due to its high accuracy and exceptional speed. It is also considered an extension of the Gradient Boosting algorithm, but it differs in several key aspects. One of the most notable differences is that the addition of weak learners does not occur sequentially, but rather through a multi-threaded approach, which allows for optimal utilization of CPU cores and results in significantly enhanced performance and speed.In addition, XGBoost features a sparse-aware implementation, meaning it can automatically handle missing values without requiring manual preprocessing. It also employs a block structure that supports the parallelization of tree construction, and it enables continued training, allowing a previously trained model to be further improved using new data. It is worth noting that XGBoost has demonstrated clear superiority in handling structured or tabular datasets, particularly in classification, regression, and predictive modeling tasks. (Santhanam, et al.2016).

$$\hat{y}_i = \sum_{k=L}^{k} f_k(x_i), f_k \in F$$

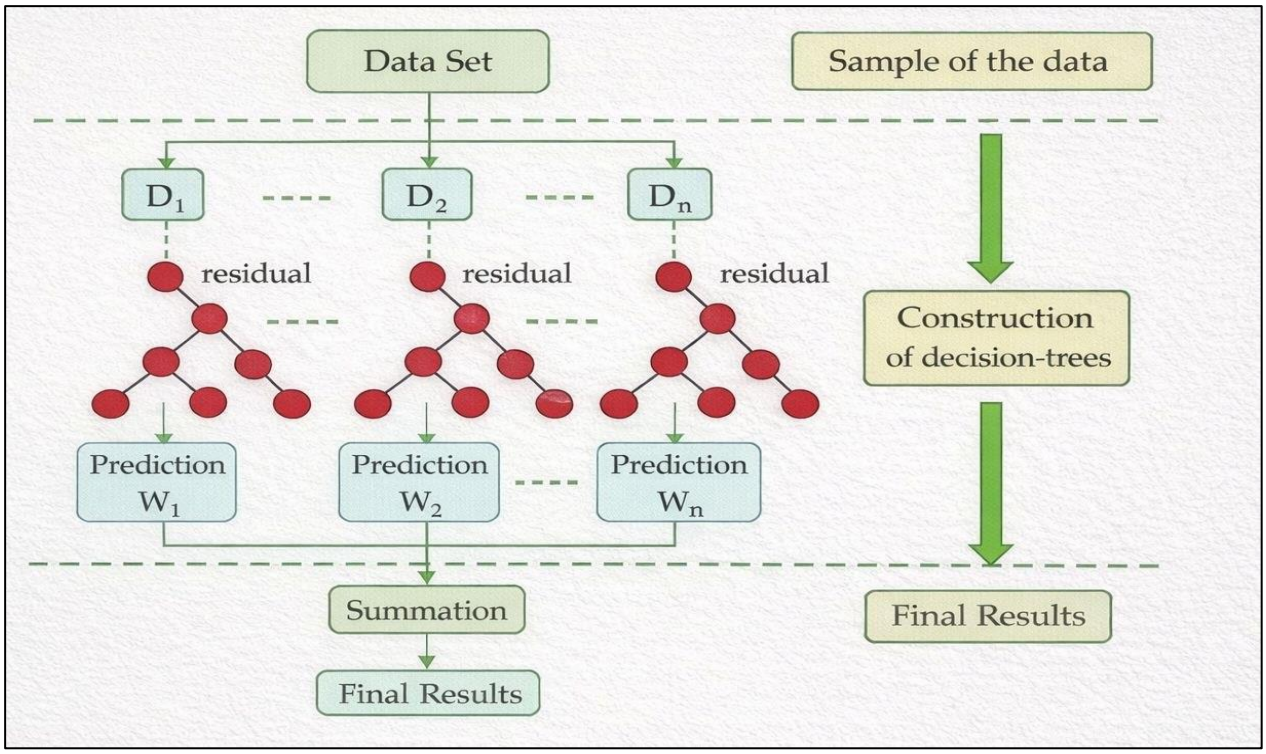


Figure 7. Xboost model (https://www.researchgate.net)

1.2.4 Unsupervised learning

Unlike supervised learning, where each training example is paired with an output label, unsupervised learning is concerned with understanding the hidden structure, relationships, or patterns that exist naturally within the data itself. The model attempts to explore the dataset and discover groupings (clusters), latent variables, or underlying features that are not immediately obvious. This process is entirely data-driven, meaning the algorithm must rely solely on the input data to make sense of its structure, without any external guidance.In unsupervised learning, the model attempts to learn the inherent structure of the data without any labels or supervision. It aims to find patterns, groupings, or features that describe the data. (Benjamin, 2025).

1.2.4.1 Unsupervised learning algorithms

1.2.4.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction, aiming to reduce the number of variables in the data while preserving as much variance as possible. The key idea is to transform the data into a new coordinate system where the axes (called principal components) capture the maximum variance in the data.

The first principal component captures the largest variance, the second captures the second largest, and so on. PCA is widely used for:

a) **Reducing dimensionality:** Simplifying data without losing essential information.
b) **Data visualization:** Making it easier to visualize complex, high-dimensional data.
c) **Noise reduction:** By removing less important components, PCA helps in filtering out noise and focusing on the critical features of the data.
d) **Preprocessing for other algorithms:** PCA can be used as a preprocessing step to make other algorithms more efficient by reducing the input dimensionality.

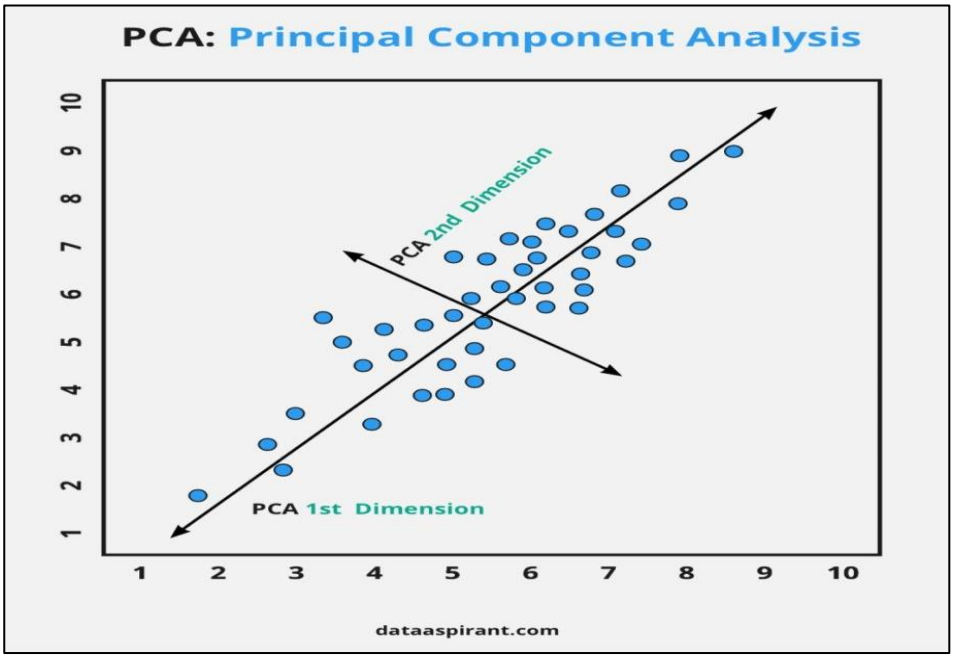


Figure 8. PCA model *(https://dataaspirant.com)*

1.2.4.1.2 Clustering

Clustering is the task of grouping data points into clusters such that data points within a cluster are more similar to each other than to those in other clusters. Various clustering algorithms are commonly used in unsupervised learning, including:

a) **K-means:** This algorithm partitions the data into K clusters based on the similarity (distance) between data points and the cluster centroids. It is widely used due to its simplicity and efficiency but works best with spherical clusters.
b) **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** DBSCAN groups data points based on their density, meaning points that are close to

each other in dense regions are grouped together, while noise or outliers are treated separately.

c) **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):** BIRCH is a clustering algorithm designed to handle large datasets by organizing data hierarchically and reducing the computational burden of clustering. Clustering algorithms are used in :

- **Customer segmentation:** Grouping customers based on their buying behavior.
- **Anomaly detection:** Identifying outliers or unusual patterns in the data.
- **Data compression:** Reducing the data size by grouping similar data points together.
- **Pattern recognition:** Detecting patterns in data that can be used for predictions for further analysis. (Benjamin, et al 2025).

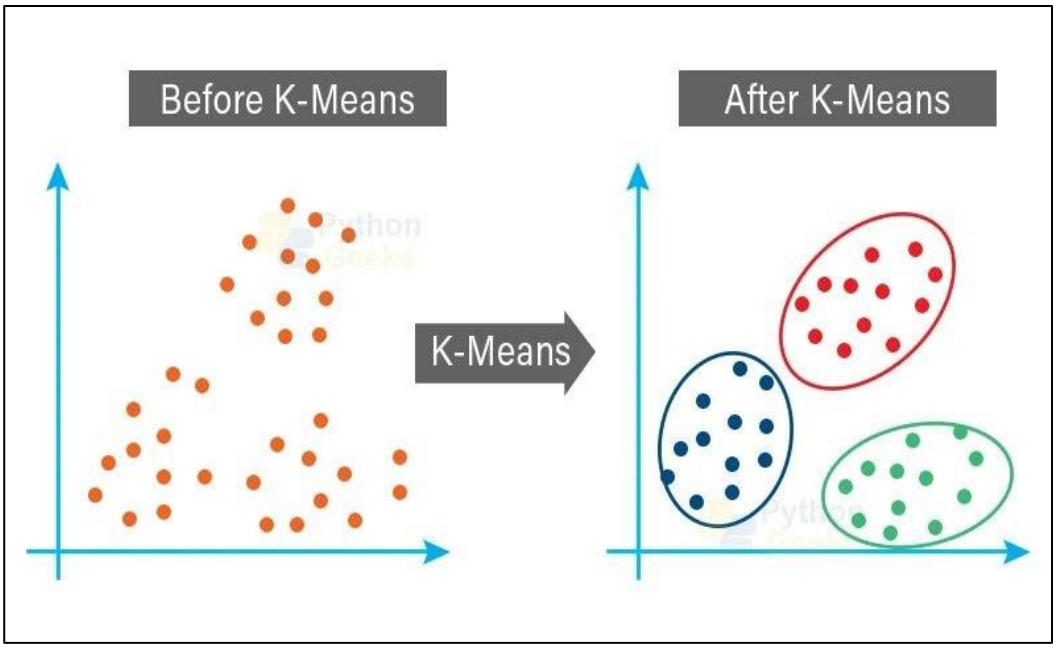


Figure 9. clustering (k- means) model (https://media.licdn.com)

1.3 Deep Learning

Deep learning is a subset of machine learning that uses deep neural networks networks with many layers to automatically learn and extract data representations at multiple levels of abstraction. These layers help the model gradually understand complex features from raw data, enabling it to perform tasks like speech recognition, image classification, and language translation with exceptional accuracy. The deeper the

network, the more abstract the representations it can form, making it a highly effective approach for handling large-scale, unstructured data. (Le Cun; et. al .2015).

Before diving into the different algorithms used in deep learning for various applications, we begin by defining the core model of deep learning “the neural network”.

1.3.1 The Structure of a biological Neuron

The structure of a neuron consists of:

a) **Cell Membrane:** The cell membrane of the neuron separates the internal and external environments of the neuron. It contains ion channels that help transmit electrical signals across the neuron.
b) **Dendrites:** Dendrites are branched extensions that receive electrical signals from other neurons or sensory receptors. They are tree-like structures that capture neural input and convert it into electrical signals.
c) **Cell Body (Soma):** The cell body contains the neuron's nucleus and is responsible for maintaining the neuron's functions and generating energy. It processes the signals received from the dendrites.
d) **Axon:** The axon is a long tube-like structure that transmits electrical signals from the cell body to the neuron’s terminal, where the signal is transmitted to other neurons or cells.
e) **Myelin Sheath:** A fatty layer wrapped around the axon that helps increase the speed of electrical signal transmission. This sheath acts as an electrical insulator and improves the efficiency of the neural signal.
f) **Nodes of Ranvier:** These are the gaps between myelinated sections of the axon where electrical signals can jump through open ion channels, enhancing the speed of signal conduction along the axon.
g) **Synapse:** The synapse is the junction where one neuron communicates with another neuron or target cells, such as muscle cells or glands. The neural signal is transmitted across this point using chemical messengers (neurotransmitters) that are released from the axon terminal and affect neighboring cells.
h) **Neurotransmitters:** These are chemical substances that transmit the neural signal across the synapse to the next neuron or target cell. These substances include serotonin, dopamine, acetylcholine, and others. (Marder; et .al .2011).

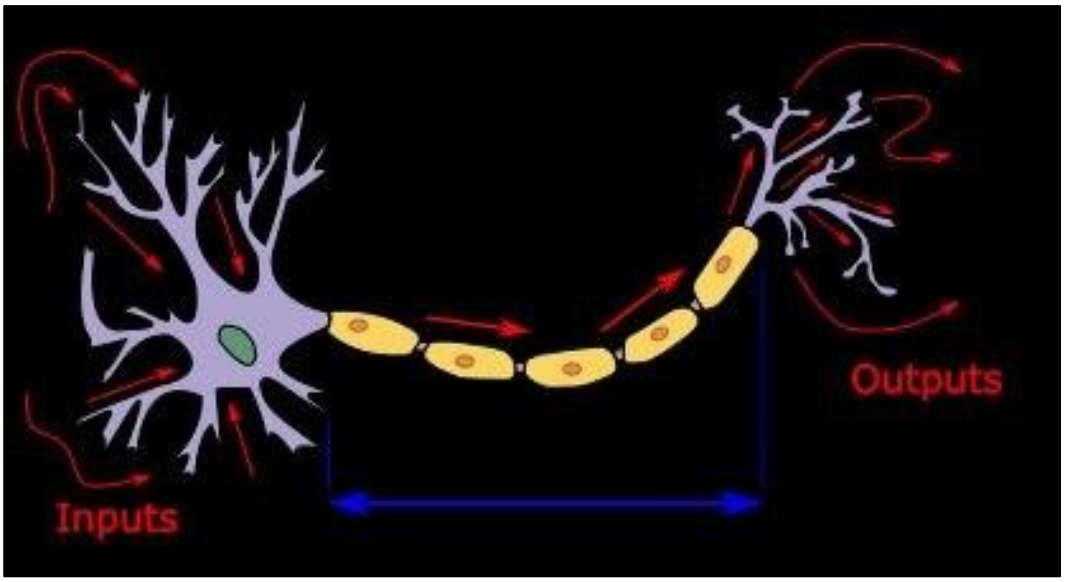


Figure 10. *a typical model of biological neural (https://media.springernature.com)*

1.3.2 Artificial neural network

Artificial Neural Networks (ANNs) represent a branch of artificial intelligence designed to simulate intelligent behavior by mimicking biological neural networks. They have evolved from niche algorithms into fundamental tools for data analysis and pattern recognition, driven by significant advancements in computational power. Initially met with skepticism in scientific fields, ANNs have proven their effectiveness across various disciplines, particularly in analytical chemistry and biology. This shift in perception is reflected in the growing number of research publications employing ANNs, especially in chemistry, underscoring their increasing acceptance and integration into scientific research. (Zou; et. al. 2009).

1.3.3 Principle of the Artificial Neuron

Artificial Neural Networks (ANNs) operate based on a computational model inspired by the structure and function of biological neural networks in the brain. The fundamental concept is that ANNs learn from data by adjusting internal parameters (called weights) to recognize patterns and make predictions or classifications. Here's a breakdown of how ANNs work, layer by layer:

a) **Input Layer:** This layer serves as the entry point of the neural network, receiving the raw data that will be analyzed. It consists of several nodes (neurons), each representing a single input feature or variable.

b) **Hidden Layers:** These layers are responsible for the core processing of the network, were information flows through weighted connections between neurons. A network may contain one or more hidden layers depending on the complexity of the task. Each neuron applies an activation function (such as ReLU, Sigmoid, or Tanh) to the weighted sum of its inputs, introducing non-linearity and enabling the network to model complex patterns. During training, the network optimizes the weights using algorithms like backpropagation by minimizing errors through gradient descent.

c) **Output Layer:** This layer produces the outcome of the neural network, which could be a prediction or a classification. The structure of this layer varies according to the nature of the task, with the number of neurons and activation functions tailored to generate the appropriate outputs.

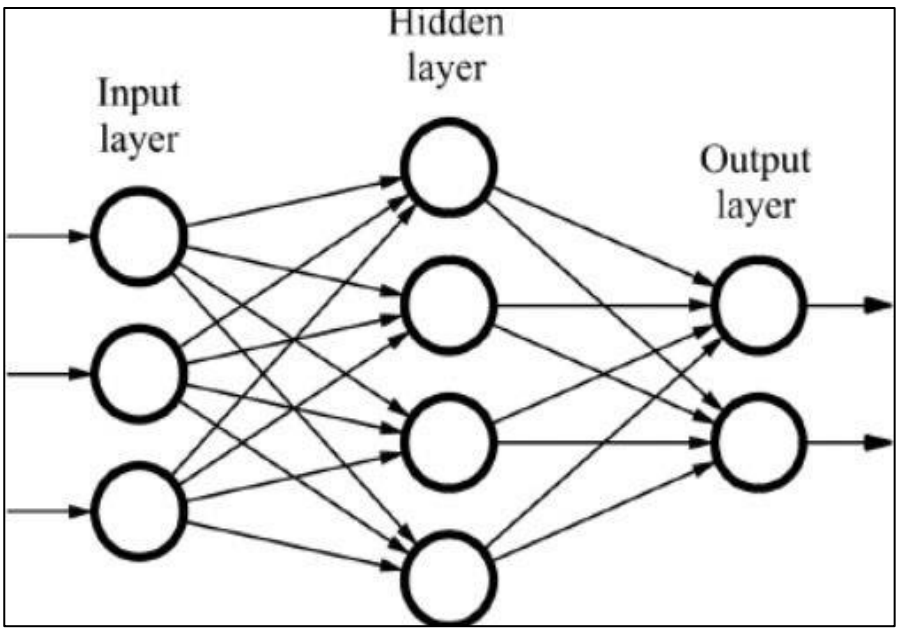


Figure 11. Artificial neural network (https://media.geeksforgeeks.org)

2 Applications of Artificial Intelligent in Petroleum Engineering

2.1 Exploration

The use of AI in oil exploration has numerous benefits. It can improve exploration accuracy, reduce exploration costs, increase safety, and optimize drilling operations. To use AI in oil exploration, companies need to collect and analyze large amounts of data and use AI algorithms to identify patterns and make predictions. With the continued

advancement of AI technology, we can expect to see even more innovative applications in the future.

Benefits of Using AI in Reduction **in Dry Well Exploration:** Modern AI-based techniques have enabled upstream oil and gas companies to ascertain geological features like rock formation for ensuring they don't waste time exploring dry wells. In addition, geologists are utilizing well-log data to develop ML models with existing and new seismic profiles, and it's helping them make an educated guess regarding the location of potential oil and gas reservoirs (Bajaj, 2021).

2.2 Seismic Image and Surveying

Exploring oil and gas reservoirs requires 3D imaging of the field and processing the data acquired from petrophysical and geophysical studies, including seismic surveying at the scale of the reservoir. AI reduced the time taken to process these 3D images, an otherwise time-intensive process.

AI can be used to analyze these images and identify patterns that are difficult for humans to detect. For example, AI can identify subtle changes in rock formations that may indicate the presence of oil or gas.

Furthermore, the 3D images (or seismic cubes) are studied and segmented by experts, and the whole process can take more than a year to conduct accurate seismic study. Using modern deep learning-based pattern recognition techniques makes it possible to accelerate the interpretation process by a factor of 10-1000 (Bajaj, 2021).

2.3 Reservoir Characterization

Reservoir characterization is a crucial aspect of petroleum engineering, as it involves understanding the properties of underground reservoirs and identifying the most effective methods for extracting oil and gas. Artificial Intelligence (AI) is increasingly being used in a reservoir of characterization to improve efficiency, reduce costs, and increase accuracy.

AI is applied in reservoir characterization through several techniques, such as:

- Predictive Analytics AI algorithms can analyze large amounts of geological and production data to predict future reservoir behavior. This helps with optimizing production and minimizing the risk of costly errors.

- Image Recognition AI can analyze seismic images and well logs to identify rock properties and fluid content in the reservoir. This helps in accurately characterizing the reservoir and optimizing production.
- AI algorithms can analyze seismic data to identify potential drilling locations. This can help engineers to optimize drilling and completion, improve recovery rates, and reduce costs.
- AI being used in reservoir characterization is the use of machine learning algorithms to analyze well logs. Machine learning algorithms can identify patterns in well logs that are not visible to the human eye, enabling them to predict oil and gas reservoir properties such as porosity, permeability, water saturation, well-bore stability, and identification of lithofacies with greater accuracy (Anifowose et al., 2011).
- Neural networks can identify subtle patterns in seismic data that are difficult for humans to detect, enabling them to identify potential drilling locations with greater accuracy (Raza et al., 2020).
- **Data Integration:** AI can integrate data from various sources such as well logs, seismic data, and production data to create a comprehensive model of the reservoir. This helps in better understanding the reservoir behavior and optimizing production. ultimately leading to increased profitability and sustainable resource management.
- Due to the decrease in commodity prices in a constantly dynamic environment, there has been a constant urge to maximize benefits and attain value from limited resources. Traditional empirical and numerical simulation techniques have failed to provide comprehensive optimized solutions in a relatively short time. Coupled with the immense volumes of data generated daily, a solution to tackle industry challenges became imminent. Various expert opinion fraught with bias has posed extra challenges to obtain timely cost-effective solutions. Data Analytics has provided substantial contributions in several sectors. However, its full potential value has not been captured in the Oil and Gas industry yet (KFUPM, 2021).
- AI can help in automating the process of formation evaluation and reducing the time and cost required to characterize the reservoir. This can lead to more accurate reservoir models and better decision-making in the oil and gas industry.

3 Evolution of Artificial Intelligence Techniques for Real-Time Petrophysical Parameter Prediction

3.1 Early Development of AI Applications in Petrophysical Estimation (2019–2021)

During the early stage of research, Artificial Intelligence techniques started to be investigated for the estimation of petrophysical parameters in real time, particularly porosity and permeability. These two parameters are considered fundamental in reservoir characterization because the conventional Flow Zone Indicator (FZI) is mathematically derived from the relationship between effective porosity and permeability. Consequently, the first AI-based studies focusing on porosity and permeability prediction established the scientific basis for later FZI-oriented models.

At this stage, most studies relied primarily on Artificial Neural Networks (ANN) and Deep Neural Networks (DNN). Zolotukhin et al. (2019) and Arigbe et al. (2019) demonstrated that neural-network-based models could successfully estimate permeability using drilling-related data. Later, in 2020, additional approaches such as Genetic Algorithms (GA) and Multilayer Long Short-Term Memory (MLSTM) networks were introduced for predicting porosity and permeability from drilling information, as reported by Al-AbdulJabbar et al. (2020), Wang et al. (2020), and Chen et al. (2020).

By 2021, research efforts expanded toward more diversified Machine Learning methods including Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Neural Networks (NN), and XGBoost Regressor (XGBR). These approaches significantly improved the estimation accuracy of porosity and permeability using drilling and MWD data. Studies conducted by Tian et al. (2021), Gamal et al. (2021), Farouk et al. (2021), Tembely et al. (2021), and Al-Sabaa et al. (2021) confirmed the effectiveness of these models for real-time reservoir-property estimation.

Although some studies, such as Otchere et al. (2021), incorporated water saturation prediction.

3.2 Expansion of Machine Learning Approaches and Performance Improvement (2022-2023)

Between 2022 and 2023, research activity increased considerably, accompanied by broader algorithm diversification and better predictive performance. During this period,

researchers started integrating more advanced learning models capable of handling nonlinear relationships between drilling variables and reservoir properties.

Ameur-Zaimeche et al. (2022) applied several algorithms including Extreme Learning Machine (ELM), Random Forest Regressor (RFR), Multilayer Perceptron Neural Network (MLPNN), General Regression Neural Network (GRNN), and Support Vector Regression (SVR) for porosity prediction using mud gas and drilling data. These studies are highly relevant to the present thesis because mud logging data are continuously available during drilling operations and can indirectly capture lithological and reservoir-quality variations.

Simultaneously, Gradient Boosting and XGBoost algorithms became increasingly popular due to their strong predictive capabilities. Mohammadian et al. (2022) used XGBoost for porosity estimation, while Qalandari et al. (2022) employed ANN and XGB models for permeability prediction from integrated logging and drilling datasets.

Other investigations, such as those by Elmorsy et al. (2022), Nande et al. (2022), and Ezekiel et al. (2022), concentrated on permeability prediction using ANN, RF, and NN algorithms. Since permeability represents one of the principal variables involved in conventional FZI determination, these works provided important methodological support for flow-related reservoir characterization using drilling parameters.

In 2023, studies continued to explore algorithms such as KNN, RF, SVR, XGB, MLR, ANN, and SVM for real-time petrophysical estimation. Ouladmansour et al. (2023) focused on porosity prediction from drilling data, whereas Sharma et al. (2023) integrated Gamma Ray and drilling information to estimate porosity and bulk density. These findings confirmed that drilling-derived parameters can contain valuable indirect information about reservoir quality even when core measurements are unavailable.

3.3 Recent Advances in Real-Time Reservoir Characterization (2024–Present)

Recent research trends have increasingly focused on integrating Machine Learning with real-time drilling data and explainable AI techniques to improve reservoir characterization accuracy. Studies conducted by Hassaan et al. (2024) employed RF, Support Vector Machine, DT, and Gradient Boosting Regressors to estimate porosity and permeability directly from drilling data. Similarly, Anifowose et al. (2024) investigated the use of ANN, DT, RF, and SVM models with mud gas datasets for porosity prediction.

These studies are particularly important for the current work because they demonstrate the feasibility of utilizing mud logging and drilling parameters for continuous real-time evaluation of reservoir properties.

Regarding the Flow Zone Indicator itself, the conceptual foundation still originates from the Hydraulic Flow Unit methodology introduced by Amaefule et al. (1993), where FZI is derived from the relationship between Reservoir Quality Index (RQI) and normalized porosity. Therefore, AI-based porosity and permeability prediction studies can be considered the direct scientific foundation for modern FZI prediction approaches.

Conclusion

Machine learning has emerged as one of the most powerful modern tools that has brought about a qualitative transformation in the fields of Earth sciences and remote sensing. Its strength lies in its ability to process large volumes of diverse and complex data, and to infer nonlinear relationships without the need for explicit physical models. Recent studies have shown that machine learning algorithms such as artificial neural networks and support vector machines have been successfully applied in various areas including bias correction, data retrieval, code acceleration, the prediction of trace gas concentrations, surface product analysis, vegetation monitoring, and more recently, ocean monitoring. (Salcedo-Sanz et al., 2020). Machine learning techniques are characterized by their generalization capability through training on multi-source datasets. This allows them to effectively handle environmental data challenges, which are often incomplete or vary in precision. Research has demonstrated that integrating machine learning with remote sensing data improves the accuracy of predictive models and enhances our understanding of complex environmental systems. (Binetti et al., 2024) With continuous advancements in computing technologies and the increasing availability of geophysical and environmental data, machine learning is expected to play an increasingly important role in developing more accurate analytical tools, improving the effectiveness of forecasts, and supporting scientific decision-making based on data. Moreover, it is anticipated to accelerate innovation and offer smart, efficient solutions to current environmental challenges this enhancing our understanding of natural systems and supporting more sustainable and effective resource management efforts. (Peng et al., 2023).

Chapter 02
Geological and Reservoir Overview of the Study Area

Introduction

The Berkine Basin occupies an important position among the hydrocarbon basins of the Algerian Sahara. Its petroleum interest is mainly related to its large areal extent, its long geological evolution, and the presence of several proven source rocks and reservoir levels. For this reason, understanding the regional geology of the basin is a necessary step before discussing the characteristics of any local field, particularly when the study deals with reservoir quality and petrophysical parameters.

This chapter first gives a general geological description of the Berkine Basin, including its geographical position, structural framework, stratigraphic succession, and petroleum system. The main elements controlling hydrocarbon accumulation, such as source rocks, reservoir formations, seals, migration pathways, and traps, are also discussed in order to place the study area within its regional context.

The second part of the chapter is devoted to the Sif Fatima Field, which represents the study area of this work. The discussion focuses on its location, local stratigraphy, structural setting, and especially the TAGI reservoir. This reservoir is mainly composed of fluvial deposits, where the distribution of sandstone channels and clay-rich intervals strongly controls reservoir quality, porosity, and permeability.

Thus, this geological background is important for the following chapters, especially for interpreting petrophysical data, evaluating permeability behavior, and understanding the reservoir quality variations within the Sif Fatima Field.

1 Regional Geology of the Berkine Basin

The Berkine Basin represents an important hydrocarbon province in the Algerian Sahara. It is primarily an oil and gas basin (Boukhenak et al., 2024). Geological and geophysical studies carried out in the region (Berkine) have mainly focused on Sif Fatama.

1.1 Geographical Setting

The Berkine Basin (formerly known as the Ghadames Basin) is located in the eastern Erg of the Algerian Sahara, between latitudes 29° and 34° North and longitudes 5° and 10° East. It is bounded to the east by the Syrte Basin between Tunisia and Libya, characterized by a series of NW–SE trending faults. To the south, it is limited by the Illizi Basin; to the west by the Amguid El Biod–Hassi Messaoud high; and to the north by the Ain-Roumana swell and the Dahar arch **(Ameur Zaimeche, 2014)** It covers an area of nearly 350,000 km².

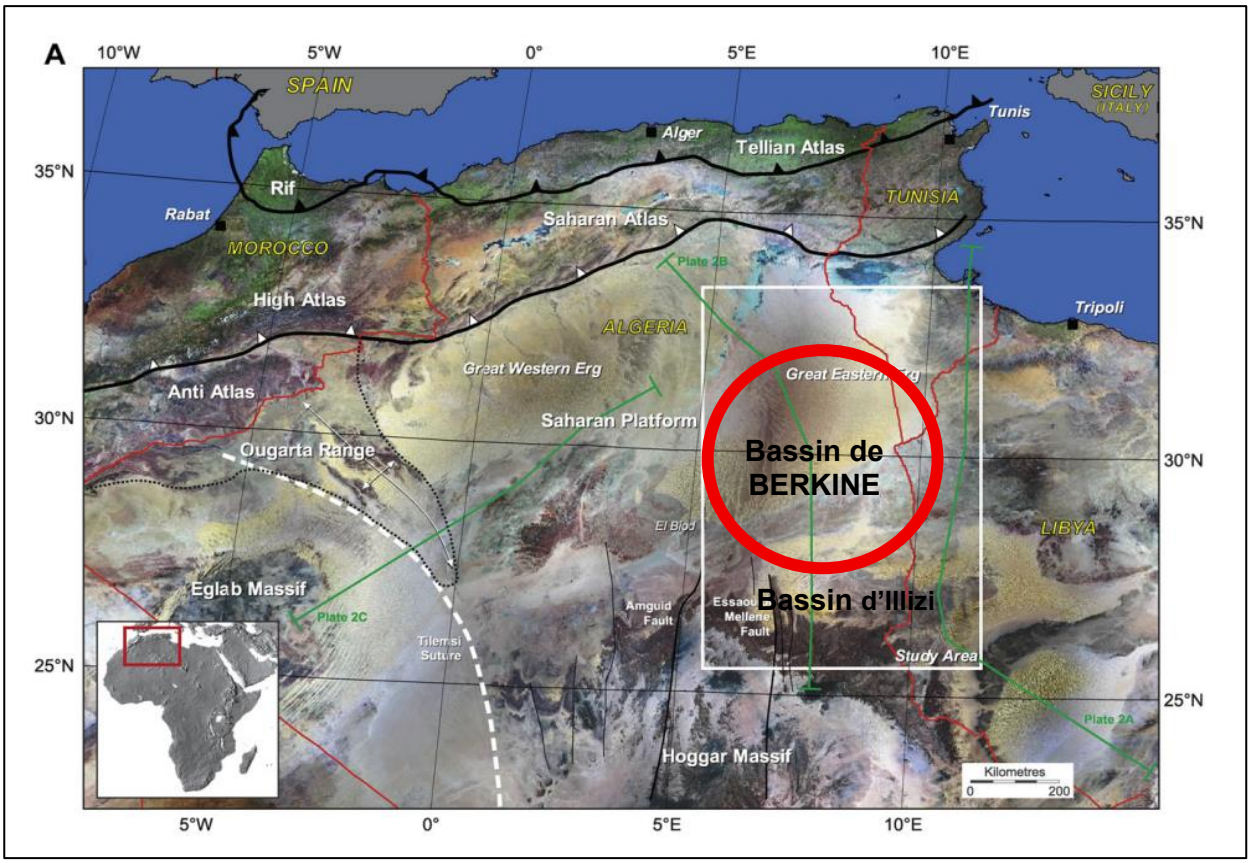


Figure 12. *Geographical Setting of the Berkine Basin (S. Galeazzi et al., 2010), modified*

1.2 Regional Geological Framework

From a geological point of view, the Berkine Basin is classified among intracratonic basins (WEC, 2007). It is located in the northeastern part of the Saharan platform and corresponds to a platform basin resting on a basement of infracambrian age.

The basin covers a total area of approximately 300,000 km², of which about 103,000 km² lie within Algerian territory. The basement, reached at depths of up to 7,000 m (absolute elevation), is mainly composed of crystalline, metamorphic, and volcanic rocks. Above this basement lies an unconformable sedimentary cover composed of formations ranging in age from the Cambrian to the present day. **(Ameur Zaimeche, 2014)**

2 Stratigraphic Framework

The Berkine Basin is characterized by a thick sedimentary cover exceeding 6,000 meters in thickness, ranging from the Paleozoic era to the present. This sedimentary sequence rests on a major unconformity known as the infra-Tassilian unconformity, and is mainly composed of Mesozoic deposits overlying Paleozoic formations. (Ameur Zaimeche, 2014; Omiri & Dahmani, 2021).

The stratigraphic succession of the Berkine Basin can be summarized from bottom to top as follows :

2.1 Basement (The Basement Complex)

It consists of Precambrian rocks (Eburnian and Proterozoic), mainly igneous rocks (such as granite) and metamorphic rocks (such as gneiss and schist). Paleozoic formations rest on this basement in an unconformable manner.

2.2 Paleozoic Era (Le Paléozoïque)

It is composed of:

2.2.1 Cambrian :

Fluvial deposits consisting of sandstone (grès) and quartzite with conglomeratic interbeds, with an average thickness of about 300 m.

2.2.2 Ordovician :

Clastic sediments divided into several lithological units from oldest to youngest: El Gassi clays, El Atchane sandstone, Hamra quartzites, Ouargla sandstone, Azzel clays, Oued Saret sandstone, micro-conglomeratic clays, and Ramade sandstone.

2.2.3 Silurian :

Fine marine clay deposits mainly composed of black organic-rich shales containing fossils (graptolitic clays), overlain by argillaceous sandstone. The black shales represent an important stratigraphic marker across the Saharan platform.

2.2.4 Devonian :

Divided into three parts:

- Lower Devonian: clay-sandstone formations
- Middle Devonian: clay deposits with carbonate interbeds
- Upper Devonian: clay with marl, limestone, and dolomite intercalations, overlain by sandstone and silt layers

2.2.5 Carboniferous :

The lower part consists of alternating clay, silt, sandstone, and carbonate layers, while the upper part is characterized by sandstone–clay alternations with carbonates in the central interval.

2.3 Mesozoic Era (Le Mésozoïque)

2.3.1 Triassic: Divided into :

a) **TAC (clay-carbonate Triassic):** brown-red to grey-green clays with hard clay and siliceous interbeds
b) **TAGI (lower clay-sand Triassic):** sandstone with clay intercalations
c) **TAGS (upper clay-sand Triassic):** compact sandstone with clay interbeds
d) **Salt-bearing Triassic (Trias S4):** containing layers of salt and anhydrite

2.3.2 Jurassic :

Marine, continental, and lacustrine deposits, characterized at the base by a dolomitic level.

2.3.3 Cretaceous :

Average thickness of about 1,250 m. It consists of alternating sandstone, clay, dolomite, limestone, anhydrite, gypsum, and salt, becoming mainly carbonate toward the top.

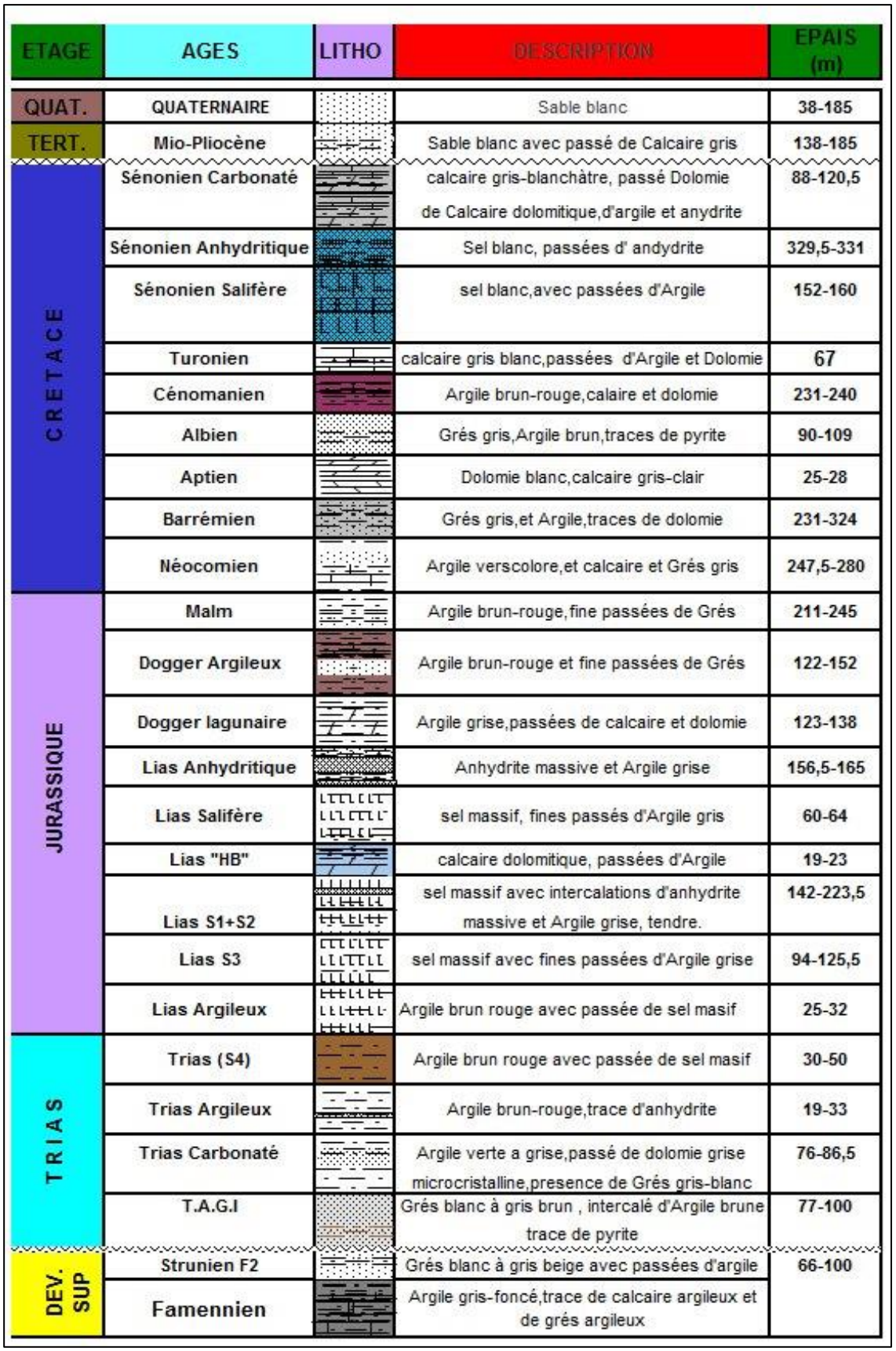
2.4 Cenozoic Era (Le Cénozoïque)

2.4.1 Mio-Pliocene:

White to translucent and yellowish sandy deposits with sandy clay interbeds and some argillaceous limestone layers.

2.4.2 Quaternary:

Represented by dune sands (sables dunaires).



ETAGE	AGES	LITHO	DESCRIPTION	EPAIS (m)
QUAT.	QUATERNAIRE		Sable blanc	38-185
TERT.	Mio-Pliocène		Sable blanc avec passé de Calcaire gris	138-185
CRETACE	Sénonien Carbonaté		calcaire gris-blanchâtre, passé Dolomie de Calcaire dolomitique,d'argile et anydrite	88-120,5
	Sénonien Anhydritique		Sel blanc, passées d' andydrite	329,5-331
	Sénonien Salifère		sel blanc,avec passées d'Argile	152-160
	Turonien		calcaire gris blanc,passées d'Argile et Dolomie	67
	Cénomanien		Argile brun-rouge,calaire et dolomie	231-240
	Albien		Grés gris,Argile brun,traces de pyrite	90-109
	Aptien		Dolomie blanc,calcaire gris-clair	25-28
	Barrémien		Grés gris,et Argile,traces de dolomie	231-324
	Néocomien		Argile verscolore,et calcaire et Grés gris	247,5-280
JURASSIQUE	Malm		Argile brun-rouge,fine passées de Grés	211-245
	Dogger Argileux		Argile brun-rouge et fine passées de Grés	122-152
	Dogger lagunaire		Argile grise,passées de calcaire et dolomie	123-138
	Lias Anhydritique		Anhydrite massive et Argile grise	156,5-165
	Lias Salifère		sel massif, fines passés d'Argile gris	60-64
	Lias "HB"		calcaire dolomitique, passées d'Argile	19-23
	Lias S1+S2		sel massif avec intercalations d'anhydrite massive et Argile grise, tendre.	142-223,5
	Lias S3		sel massif avec fines passées d'Argile grise	94-125,5
	Lias Argileux		Argile brun rouge avec passée de sel masif	25-32
TRIAS	Trias (S4)		Argile brun rouge avec passée de sel masif	30-50
	Trias Argileux		Argile brun-rouge,trace d'anhydrite	19-33
	Trias Carbonaté		Argile verte a grise,passé de dolomie grise microcristalline,presence de Grés gris-blanc	76-86,5
	T.A.G.I		Grés blanc à gris brun , intercalé d'Argile brune trace de pyrite	77-100
DEV. SUP	Strunien F2		Grés blanc à gris beige avec passées d'argile	66-100
	Famennien		Argile gris-foncé,trace de calcaire argileux et de grés argileux	

Figure 13. Lithostratigraphic Log of the Berkine Basin (Sonatrach / Anadarko 2012).

3 Petroleum System

This system consists of the following key elements that ensure the generation, migration, and accumulation of hydrocarbons:

3.1 Source Rocks

3.1.1 Silurian Source Rock :

Composed of marine clay rich in organic matter and graptolite fossils. This rock underwent burial during the Carboniferous, entering the “oil window” (fenêtre à l'huile) and generating petroleum. It later experienced a second maturation phase during the Cretaceous, leading to gas generation (Sediri, 2017; Omiri & Dahmani, 2021).

3.1.2 Devonian / Frasnian Source Rock :

Of marine origin, it began generating oil during the Late Cretaceous. It expelled hydrocarbons toward the Strunian, Carboniferous, and Triassic reservoirs (Sediri, 2017; Omiri & Dahmani, 2021).

3.1.3 Secondary Source Rocks:

Represented by Ordovician and Carboniferous formations (Sediri, 2017).

3.2 Reservoir Rocks

The region is characterized by multiple reservoir rocks with good petrophysical properties (high porosity and permeability), including:

3.2.1 Triassic Reservoirs :

The most important is the Lower Triassic clay-sand unit (TAGI), considered a main reservoir with porosity ranging between 4% and 11%. The Upper Triassic clay-sand unit (TAGS), of fluvial origin, is also significant (Sediri, 2017; Omiri & Dahmani, 2021).

3.2.2 Silurian Clay-Sand Reservoirs (SAG) :

Mainly represented by reservoir (F6) and its subdivisions (M1, M2, A, B1, B2), with porosity ranging between 8% and 12% (Sediri, 2017).

3.2.3 Lower Devonian Reservoirs (Gedinnian and Siegenian) :

Composed of fluvial to deltaic sandstones, often containing condensate gas and light oil.

3.2.4 Carboniferous Reservoirs:

Characterized by very good porosity ranging from 8% to 17%.

3.2.5 Ordovician and Cambrian Reservoirs:

Such as the Hamra quartzites (Ordovician) and Cambrian reservoirs, whose productivity may sometimes depend on fracturing.

3.3 Seal Rocks (Cap Rocks)

To preserve hydrocarbons and prevent leakage, impermeable layers are present:

a. Seal of Triassic reservoirs (TAGI and TAGS):

ensured by clays and evaporites from the carbonate Triassic and the salt-bearing Triassic (S4).

b. Seal of Devonian, Carboniferous, and Silurian reservoirs:

provided by intraformational clay layers from the same periods.

c. Seal of Cambrian and Ordovician reservoirs:

mainly ensured by El Gassi clays.

d. Lateral seal :

faults (fault throws) contribute to lateral trapping and prevent hydrocarbon escape.

3.4 Migration

3.4.1 Direct migration :

hydrocarbons migrated from Silurian and Frasnian source rocks directly into Lower Triassic and Silurian (F6) reservoirs beneath the Hercynian unconformity.

3.4.2 Fault-assisted migration:

Upper Triassic and Ordovician reservoirs (such as Hamra) relied on fault networks as migration pathways.

3.5 Traps

Traps in the Berkine Basin are diverse and classified into :

a. Structural Traps :

Related to Hercynian and Austrian compressional phases, including simple anticlines (e.g., RKF field) and faulted anticlines (e.g., Ourhoud and Gourde El Nous).

b. Stratigraphic Traps :

Represented by pinch-outs beneath unconformities, sandstone lenses (Triassic and Carboniferous), and lateral facies changes. In the study area, this type is dominant for Silurian units (F6).

c. Mixed Traps :

Combining both structural and stratigraphic characteristics.

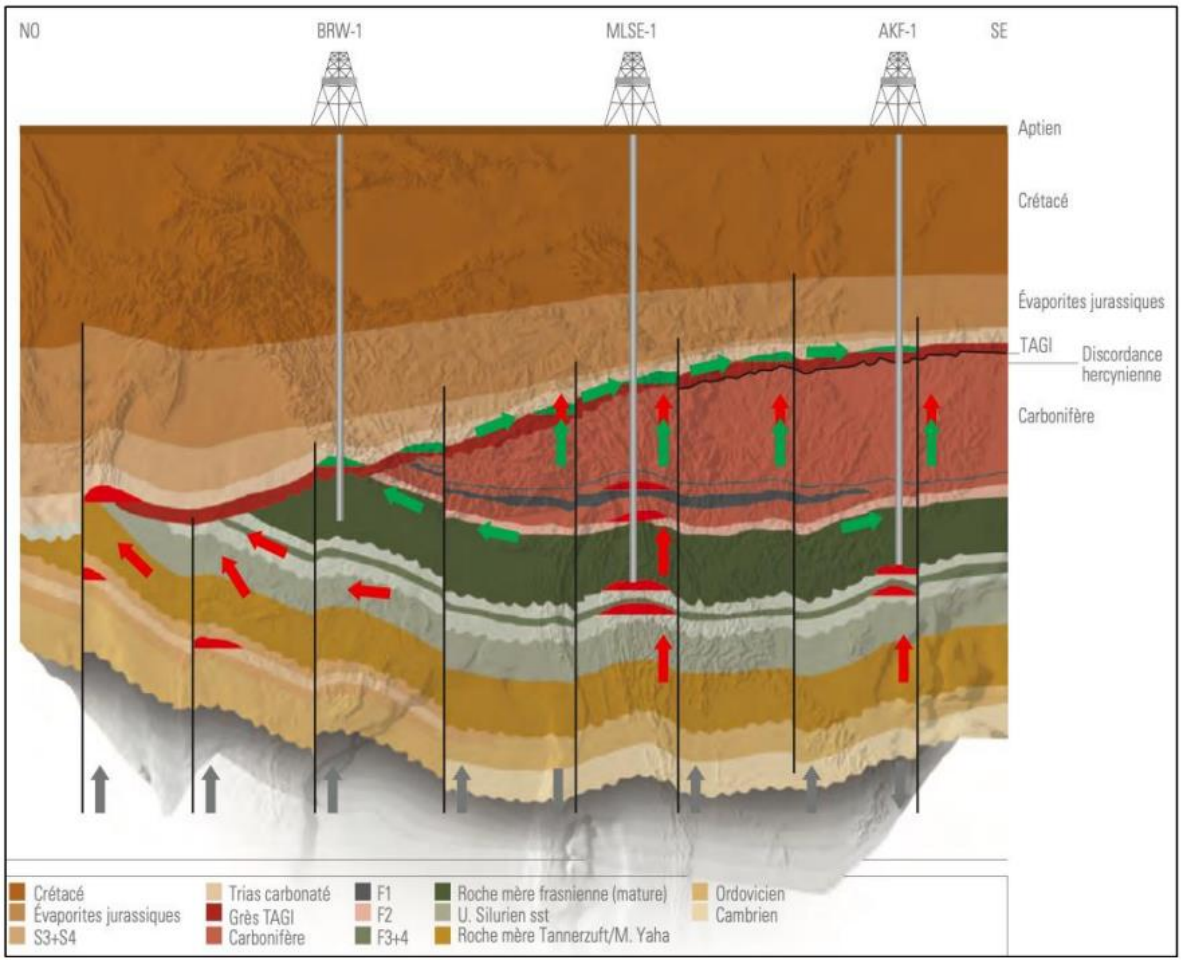


Figure 14. Petroleum System in the Berkine Basin (Galeazzi et al., 2010)

1 Local geology of the sif fatima field

The Sif Fatima Field (SF) is one of the important oil fields in the Berkine Basin It is characterized by specific geological and petrophysical features that make it an excellent case study for reservoir characterization.

3.6 Geographical and Geological Setting

The Sif Fatima Field is located in the northeastern part of the Berkine Basin This sector (402) covers an area of approximately 103 km², Geographically, the study area lies between latitudes 31° and 32° North and longitudes 8° and 9° East.

The field is situated in a region with relatively complex tectonics, where traps were formed as a result of the interaction between successive tectonic movements and sedimentation processes in continental and shallow marine environments.

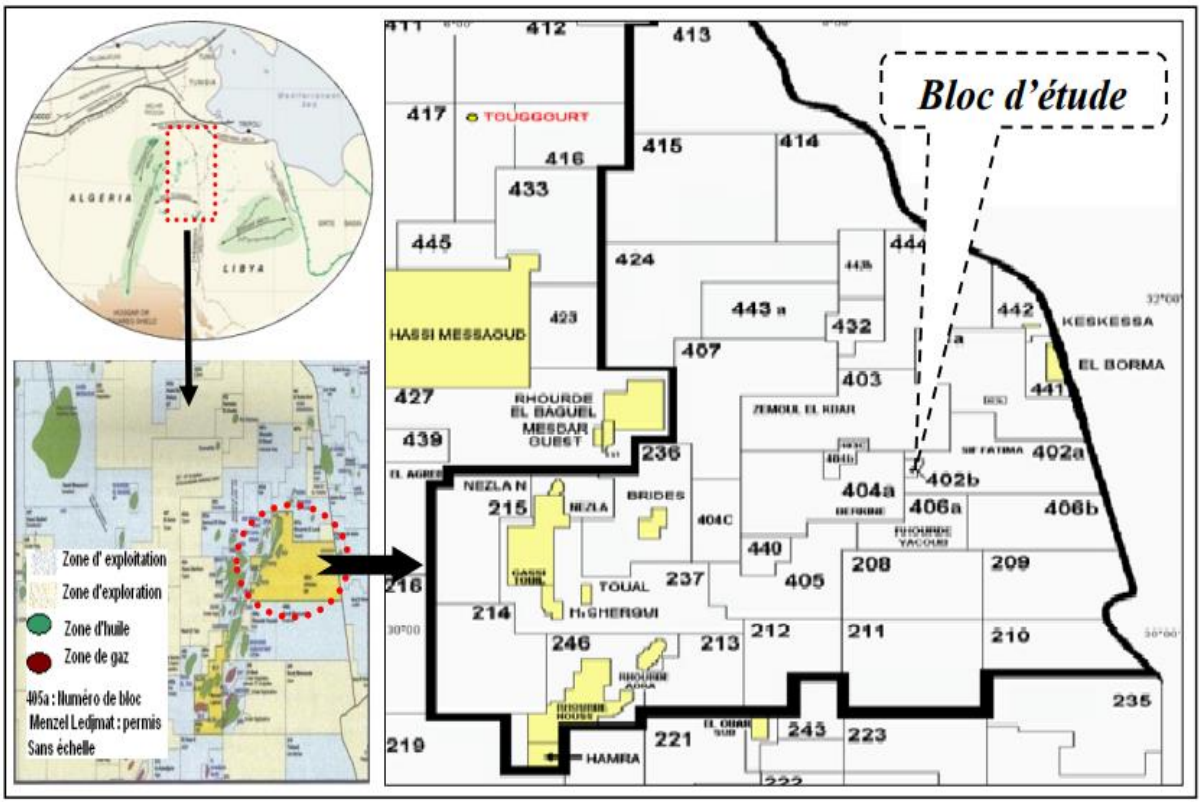


Figure 15. Location Map of the Study Block (SIF FATIMA) in the Berkine Basin **(WEC, 2007)**

3.7 Local Stratigraphy

The stratigraphic succession in the Sif Fatima Field is marked by well-developed Mesozoic formations resting on the Hercynian unconformity. The lithostratigraphic sequence from bottom to top is mainly as follows:

a. Lower Triassic Clay-Sand Formation (TAGI):

This is the main reservoir and the primary target of the field. It consists of clastic deposits of fluvial origin, represented by alternations of medium- to coarse-grained sandstone with interbedded clay layers.

b. Triassic Carbonate Formation (TAC) :

Overlies the TAGI reservoir and consists of alternating clays, marls, and carbonate/dolomite layers.

c. Salt-bearing Triassic (Trias Salifère – S4):

Represents the regional seal and an ideal cap rock. It is composed of very thick layers of halite and anhydrite.

d. Jurassic and Cretaceous :

Thick sequences of carbonates, evaporites, and marine to continental clastic deposits forming the overburden.

4 Structural Setting

Structurally, the Sif Fatima Field is the result of complex brittle tectonics :

a. Faulting :

The field is affected by a network of deep faults, mainly extensional faults formed during Triassic–Liassic rifting phases, as well as strike-slip faults.

b. Trap Type :

The trap is generally a mixed (structural–stratigraphic) trap. Structural closure is related to tilted fault blocks or gentle anticlines, while lateral facies variations (sand pinch-out or increase in clay content) contribute to stratigraphic trapping

5 Reservoir Characteristics and Facies (TAGI Reservoir)

Due to the fluvial nature of the TAGI reservoir in Sif Fatima, reservoir quality is strongly controlled by the spatial distribution of facies :

a. Fluvial Channels :

Represent the best reservoir facies, where clean sands with high porosity and permeability are deposited.

b. Flood plains :

Characterized by a high volume of clay (VCL), which reduces reservoir quality and creates permeability barriers.

c. Petrophysical Evaluation :

Accurate evaluation of this field requires detailed analysis of well logs (diagraphies) to determine variations in clay volume (VCL) and effective porosity. Rapid change in clay content play a crucial role in reserve estimation and in guiding development wells.

Conclusions

The results of this comprehensive geological and petroleum study of the Berkine Basin and the Sif Fatima Field highlight the exceptional strategic importance of this basin as one of the major energy pillars. This is clearly reflected in the role of the complex interaction between brittle tectonics and the Hercynian unconformity in forming effective structural and stratigraphic traps, supported by an integrated petroleum system that combines mature Silurian and Devonian source rocks with high-quality sandstone reservoirs such as the TAGI reservoir and the F6 unit, sealed by a regional salt cap.

The local analysis of the Sif Fatima Field revealed that the main challenge lies in the rapid variability of fluvial facies, making reservoir quality highly dependent on the accurate identification of sand channels and the avoidance of clay barriers. This ultimately emphasizes the necessity of adopting advanced methodologies in petrophysical modeling and statistical analysis of well-log data to optimize reserve exploitation and reduce technical risks in future development operations.

Chapter 03
Materials and Methods

Introduction

Reservoir characterization is a fundamental aspect of petroleum engineering that involves collecting and analyzing data to understand subsurface reservoir properties. This process is essential for making informed decisions regarding hydrocarbon production and recovery (Ertekin, 2021). It also plays a key role in reserve estimation, production optimization, and risk reduction, which are critical for economic evaluation and decision-making in the oil and gas industry (Aminzadeh et al., 2022). Additionally, it provides insights into reservoir behavior, supporting the design of well placement, production rates, and field infrastructure.

However, reservoir characterization is challenging due to reservoir heterogeneity, which reflects the variability of properties across different geological scales. To address this issue, the concept of Hydraulic Flow Units (HFUs) is used to group rocks with similar petrophysical and flow characteristics. This approach helps predict unknown reservoir properties and reduces the need for costly core measurements (Amaefule et al., 1993). HFUs are based on geological and physical flow principles and offer a more accurate representation of reservoir heterogeneity compared to traditional methods.

The Flow Zone Indicator (FZI) is closely related to HFUs and provides a quantitative measure linking microscopic properties (such as pore throat size and distribution) to macroscopic properties (such as permeability and porosity). It reflects the influence of depositional and diagenetic processes on rock characteristics (Rebelo et al., 2022).

Conventional methods rely on core measurements and well logs. While core data are highly accurate, they are expensive and limited in coverage, whereas well logs provide continuous data but may be less reliable due to interpretation uncertainties.

1 Conventional Methods

Conventional methods form the classical foundation of reservoir characterization, as they rely on integrating core data and well logs to understand petrophysical properties, especially porosity and permeability.

These methods are divided into:

- **Direct Methods:** based on laboratory measurements performed on core samples.

- **Indirect Methods:** based on the interpretation of well logs and field data.

1.1 Porosity

Porosity is the ability of a rock to store fluids, and it represents the ratio of the volume of voids within the rock to its total volume. It is usually expressed as a percentage or as a fractional value. it is defined as the pore volume (Vp) divided by the total rock volume (Vt)

The basic relationship is:

$$\Phi = \frac{\nu\rho}{\nu\tau}$$

Φ (Phi) : Porosity (-) or (%)
Vp : Pore Volume (volume of void spaces within the rock)
VT : Total Bulk Volume of the rock

1. Types of Porosity

Porosity is divided into several main types:

a) **Total Porosity:** includes all the void spaces present in the rock, whether they are connected or isolated.
b) **Effective Porosity:** includes only the interconnected pores that contribute to fluid storage and flow Effective porosity is commonly used in applied petrophysical models
c) **Primary Porosity:** original porosity formed during sediment deposition.
d) **Secondary Porosity:** porosity formed later because of dissolution, fracturing, or diagenetic processes.

1.1.1 Direct Method

Direct methods for measuring porosity rely on core samples, because the core is the only means that allows direct measurements on the actual rock in the laboratory. The plugs are taken from the core specifically to perform petrophysical measurements with high accuracy.

1.1.1.1 Helium Porosimetry

One of the most important direct methods for measuring porosity is the use of an ultra-helium porosimeter, in which the grain volume (Vs) is measured using helium

according to Boyle–Mariotte's law Porosity is then calculated from the difference between the total volume and the solid volume. This method is highly accurate because helium penetrates very small pores and does not react with the rock.

1.1.1.2 Bulk Volume Measurement

To measure the bulk volume (Vt) a mercury volumetric pump is used based on Archimedes' principle After determining the bulk volume and the solid volume, the pore volume and then the porosity can be calculated directly.

1.1.1.3 Core Plug Preparation

Before direct measurement, the plugs are washed with toluene or methanol to remove hydrocarbons, salts, and impurities, then dried in an oven at **110∘C** until a constant weight is reached. This stage is very important because any remaining material inside the pores affects the measured porosity value.

1.1.2 Indirect Methods

Indirect methods for measuring porosity rely on well logs They are faster and more comprehensive than laboratory measurements, but they remain estimated and usually need calibration with core data. The most used tools are the density log, neutron log, and sonic log. In most wells, sufficient core data are not available, so geophysical logs are used to estimate porosity indirectly.

1.1.2.1 Density Log

The density log is used to estimate porosity from the measured bulk density of the rock:

$$\boldsymbol{\phi_d = \frac{\rho ma - \rho b}{\rho ma - \rho f}}$$

Where:

ϕ_d **:** porosity calculated from the density log

ρma **:** rock matrix density

ρb **:** measured bulk density

ρf **:** fluid density in the pores

1.1.2.2 Neutron Log

This log is based on measuring the hydrogen content in the formation, and therefore it provides a good estimate of porosity, especially when the pores are saturated with water or oil. However, it can be affected by the presence of gas or shale content.

1.1.2.3 Sonic Log

The sonic log depends on the travel time of acoustic waves through the formation. One of the most widely used relationships is the Wyllie equation:

$$\boldsymbol{\phi s = \frac{\Delta tlog - \Delta tma}{\Delta tf - \Delta tma}}$$

Where :

ϕs **:** porosity derived from the sonic log

Δtlog **:** measured travel time

Δtma **:** matrix travel time

Δtf **:** fluid travel time

1.2 Permeability

Permeability is the ability of a rock to transmit a fluid through its pore space In other words, it measures how easily a fluid can flow within the rock and for this reason it is a fundamental property in evaluating reservoir quality and productivity

1.2.1 Direct Methods

1.2.1.1 Sampling Scales

Direct permeability is measured on several levels of rock samples, which differ in size and in how well they represent the rock in its original state. These levels include:

a) **Sidewall cores:** small samples taken from the borehole wall.

b) **Core plugs:** small rock plugs that are the most common standard in routine laboratory testing.

c) **Full-diameter cores:** samples that preserve the full diameter of the core.

d) **Whole-core:** complete core samples that provide the best volumetric representation of the rock

In general, the larger the sample size, the better it represents the structural and textural heterogeneity within the formation However, this usually comes at the expense of ease of preparation and testing speed

1.2.1.2 Absolute Permeability

Absolute permeability is measured when only a single phase flows through the sample, such as water or gas, and this value is considered the main reference in conventional laboratory measurements. The most commonly used method in this field is the steady-state method in which the fluid is passed through the sample after cleaning and drying, and the permeability value is then calculated by applying Darcy's law **(Kelai, 2025)**

The basic relationship is given as follows:

$$\boldsymbol{Q} = \frac{\boldsymbol{KA}}{\mu \boldsymbol{L}} \boldsymbol{\Delta P}$$

Among them :

$$\boldsymbol{K} = \frac{\boldsymbol{Q}\mu \boldsymbol{L}}{\boldsymbol{A}\boldsymbol{\Delta P}}$$

$\boldsymbol{k}$ **:** Permeability

Q: Flow rate

$\boldsymbol{A}$ **:** cross-sectional area

μ **:** fluid viscosity

$\boldsymbol{L}$ **:** sample length

$\boldsymbol{\Delta P}$ **:** pressure difference across the sample

In tight rocks, achieving a stable flow state becomes difficult. Therefore, the pulse-decay method is used, which relies on analyzing the pressure decrease over time to determine the permeability value **(Kelai, 2025)**

When the measurement is carried out using gas, the Klinkenberg correction equation is applied to correct for the gas slippage effect within very fine pores It is usually written in a simplified form as follows:

$$k_g = k_\infty(1 + \frac{b}{\rho_m})$$

This correction is applied in order to obtain a value closer to the actual liquid permeability.

1.2.1.3 Relative Permeability

Relative permeability is measured under **multiphase flow conditions,** and it expresses the behavior of each fluid phase within the rock in the presence of other competing phases in the porous medium. It is usually denoted by

$$k_{r_\alpha} = \frac{K_\alpha}{k}$$

This property is measured using two main methods:

a) **Steady-state co-flow:** where oil and water are injected together in precise proportions until a flow equilibrium is reached.
b) **Unsteady-state displacement method:** where one phase is displaced by another, with pressure and production changes monitored over time.

These measurements are essential for understanding reservoir flow mechanisms under real-world production conditions, particularly in oil recovery studies

1.2.2 Indirect Methods

1.2.2.1 Wireline-Log-Based Permeability Estimation

This method is based on estimating permeability from well log data using empirical or semi-physical relationships that link permeability to properties such as porosity**,** water saturation**,** and certain lithological indicators

1.2.2.2 Nuclear Magnetic Resonance Logging (NMR)

Nuclear Magnetic Resonance (NMR) logging is considered one of the most advanced indirect methods, because it provides information about the pore-size distribution the nature of the fluids contained in the pores and the movable free fluids Through parameters such as relaxation times and free fluid volume**,** empirical models can be developed to estimate permeability, especially in clean sandstone formations.

1.2.2.3 Geochemical Logging Tools (GLT)

Geochemical logging tools are based on identifying the elemental and mineralogical composition of the formations and then using this information in interpretative models to improve permeability estimation This method is particularly important when the flow properties of the rock are affected by mineralogical variations or by diagenetic alterations that influence the shape and connectivity of the pore system

1.2.2.4 Pressure Transient Analysis with Formation Testers (RFT, MDT)

These tests are carried out by withdrawing a small amount of fluid from the formation using tools such as RFT and MDT, then analyzing the pressure changes over time in order to evaluate the flow capacity of the layer, including its permeability This method is distinguished by providing a localized estimate under conditions that are closer to actual field conditions. It also helps differentiate between productive and non-productive layers and assess flow behavior near the borehole wall.

1.3 The flow zone indicator (FZI)

The conventional prediction of the Flow Zone Indicator (FZI) is based on the Hydraulic Flow Unit (HFU) (Amaefule et al.1993), which relies on a modified Kozeny–Carman relationship to link permeability with pore geometry and effective porosity. In this approach, permeability is expressed as a function of pore structure parameters, while the Reservoir Quality Index (RQI) and the normalized porosity (φz) are derived from core-measured porosity and permeability data. The FZI is then defined as:

$$RQI = 0,0314\sqrt{\frac{k}{\phi_e}}$$

$$\phi_z = \frac{\phi_e}{1 - \phi_e}$$

The full expanded form of FZI becomes:

$$FZI = \frac{RQI}{\phi_z}$$

FZI = Flow Zone Indicator

RQI = Reservoir Quality Index

φ_e = Effective porosity

φz = Normalized porosity (or pore volume to grain volume ratio)

A log-log plot of RQI versus φz yields straight lines of unit slope, where each line represents a distinct hydraulic unit, and the FZI corresponds to the intercept at φz = 1. Rocks within the same hydraulic unit share similar pore throat characteristics and flow behavior. Furthermore, permeability can be recalculated as a function of FZI and porosity, demonstrating that permeability is controlled not only by porosity but also by pore geometry. This method provides a more physically meaningful alternative to traditional porosity-permeability correlations, which often show large scatter and lack predictive reliability. However, the conventional FZI approach depends heavily on core data, making it costly, time-consuming, and limited in spatial coverage, especially in uncored intervals (Amaefule et al.1993).

2 Unconventional Method

Artificial Intelligence (AI) and Machine Learning (ML) have been employed more in reservoir characterisation to get beyond the drawbacks of traditional techniques. ML techniques can process large datasets, identify nonlinear patterns, and integrate multiple sources of reservoir data, and many data sources, including production, seismic, and geological data, can be integrated. While unsupervised techniques like K-Means clustering have been used for reservoir classification, a number of supervised algorithms, like as Random Forest, SVM, ANN, ANFIS, and XGBoost, have demonstrated strong performance in prediction tasks. In order to increase classification accuracy and reservoir characterization efficiency, The objective of this study is to develop a machine Learning

model a machine learning model for predicting FZI at unsampled sites using well-log data while integrating supervised and unsupervised approaches.

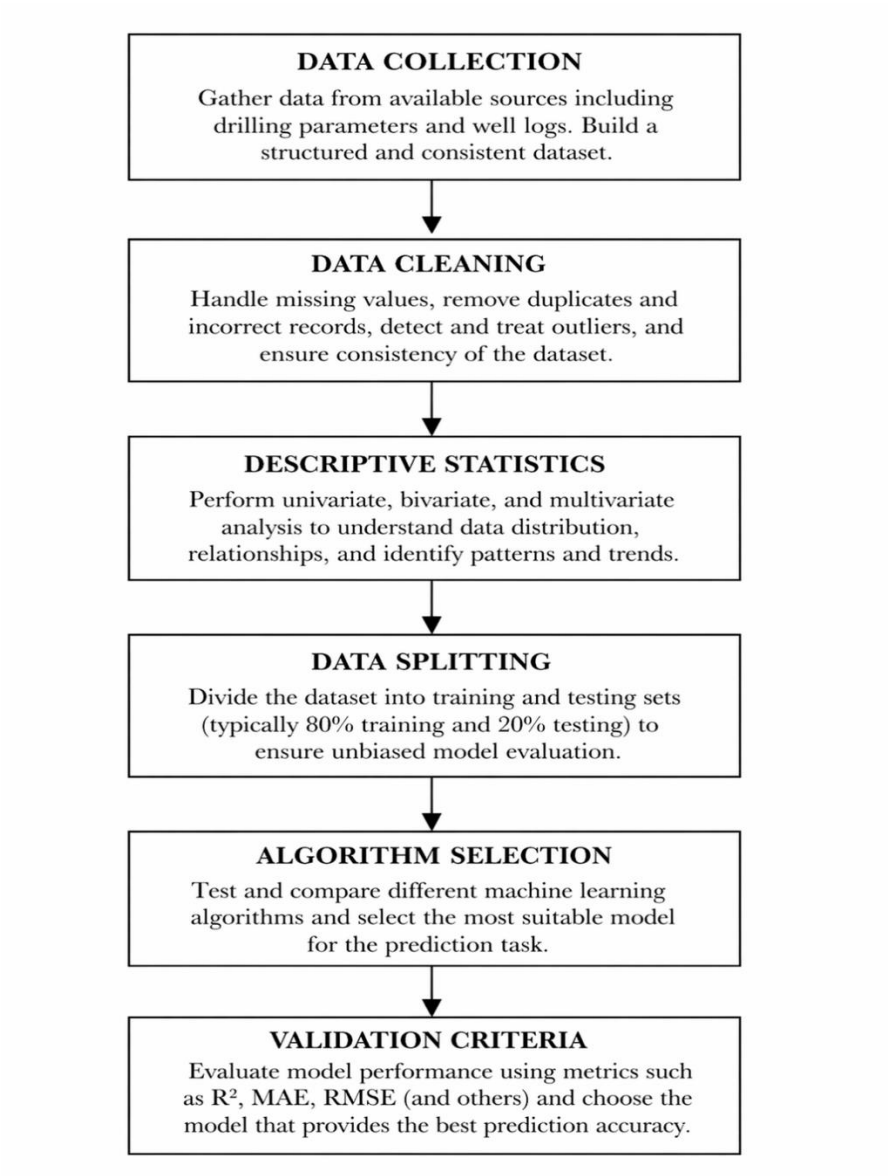


Figure 16. Methodological framework

2.1 Data Collection

The first stage consists of gathering the data required for the study from the available sources. These data may include drilling parameters, well-log measurements, or any other variables considered relevant to the prediction of the target property. At this level, the purpose is to build a structured database in which each sample contains the

input variables and the corresponding target value. The quality of this initial dataset is essential because it forms the basis of the entire predictive process.

2.2 Data Cleaning

Once the dataset is collected, it must be prepared before being used in modeling. This stage includes checking for incomplete records, removing duplicated entries, correcting inconsistent values, and identifying abnormal observations that may distort the model's behavior. Data cleaning is a necessary step because raw data are often affected by measurement errors, missing intervals, or mismatches between variables. A clean and coherent dataset improves both the stability and the credibility of the final prediction.

2.3 Descriptive Statistics

Before applying any prediction algorithm, an exploratory statistical analysis is carried out in order to better understand the characteristics of the data. This stage helps describe the distribution of variables and identify possible relationships between them. It also provides a first interpretation of the dataset and highlights trends, variability, and potential anomalies.

2.3.1 Univariate Analysis

Univariate analysis is used to examine each variable separately. It provides basic statistical indicators such as minimum, maximum, mean, and standard deviation, allowing a general understanding of the range and dispersion of the data.

2.3.2 Bivariate Analysis

Bivariate analysis focuses on the relationship between two variables, especially between each predictor and the target variable. This step is useful for identifying direct or inverse correlations and for evaluating whether a variable has a meaningful influence on the predicted property.

2.3.3 Multivariate Analysis

Multivariate analysis extends the study to several variables simultaneously. It allows a more global understanding of the interactions within the dataset and helps detect combined effects that cannot be observed through simple pairwise comparisons.

2.4 Data Splitting

At this stage, the dataset is divided into separate subsets to ensure a reliable assessment of the machine learning model. In some cases, the data are split into two subsets: a training set (80%) and a testing set (20%), where the training set is used to train the model and the testing set is used to evaluate its performance on unseen data. In other cases, the dataset is divided into three subsets: a training set (70%), a validation set (15%), and a testing set (15%). In this configuration, the training set is used to train the model, the validation set is used to tune the model and optimize its parameters during development, and the testing set is reserved for the final evaluation. Both strategies are commonly used to improve model reliability, reduce overfitting, and ensure better generalization.

2.5 Algorithm Selection

The next step consists of choosing the most suitable machine learning algorithm for the prediction task. This choice depends on the nature of the dataset, the complexity of the relationships between variables, and the expected performance of the model. Different algorithms can be tested and compared in order to determine which one provides the most satisfactory balance between prediction accuracy and robustness. The selected algorithm is then trained using the prepared dataset to learn the relationship between inputs and output.

2.6 Validation Criteria

The performance of the developed machine learning model is evaluated using statistical validation criteria that measure the agreement between the predicted values and the actual values. In this study, four main indicators are considered: the correlation coefficient (R), the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE). These criteria are used to assess the accuracy, reliability, and predictive performance of the model

The correlation coefficient (R) is used to measure the strength and direction of the linear relationship between the observed values and the predicted values. Its value ranges

from -1 to 1. A value close to 1 indicates a strong positive correlation, meaning that the predicted results are highly consistent with the actual data.

The coefficient of determination (R^2) expresses the proportion of the variance in the target variable that is explained by the model. Its value ranges from 0 to 1, and values closer to 1 indicate better model performance and stronger predictive ability. A high R^2 means that the model explains most of the variability present in the dataset.

The root mean square error (RMSE) measures the average magnitude of prediction errors by giving more weight to large errors. Lower RMSE values indicate better model performance and higher prediction accuracy.

The mean absolute error (MAE) represents the average absolute difference between predicted values and actual values. It provides a direct measure of prediction error without considering its direction Like RMSE, lower MAE values indicate a more accurate and reliable model.

Therefore, a good predictive model is generally characterized by high values of R and R^2, and low values of RMSE and MAE

Table 1. Evaluation Criteria

Metric	**Formula**
R^2 (Coefficient of Determination)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, $(0 \leq R2 < +1)$
RMSE (Root Mean Square Error)	$RMSE = \sqrt{\frac{1}{n}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, $(0 \leq RMSE < +\infty)$
MAE (Mean Absolute Error)	$MAE = \frac{1}{n}\sum_{i=1}^{n}/\ y_i - \hat{y}_i\ /$, $(0 \leq MAE < +\infty)$

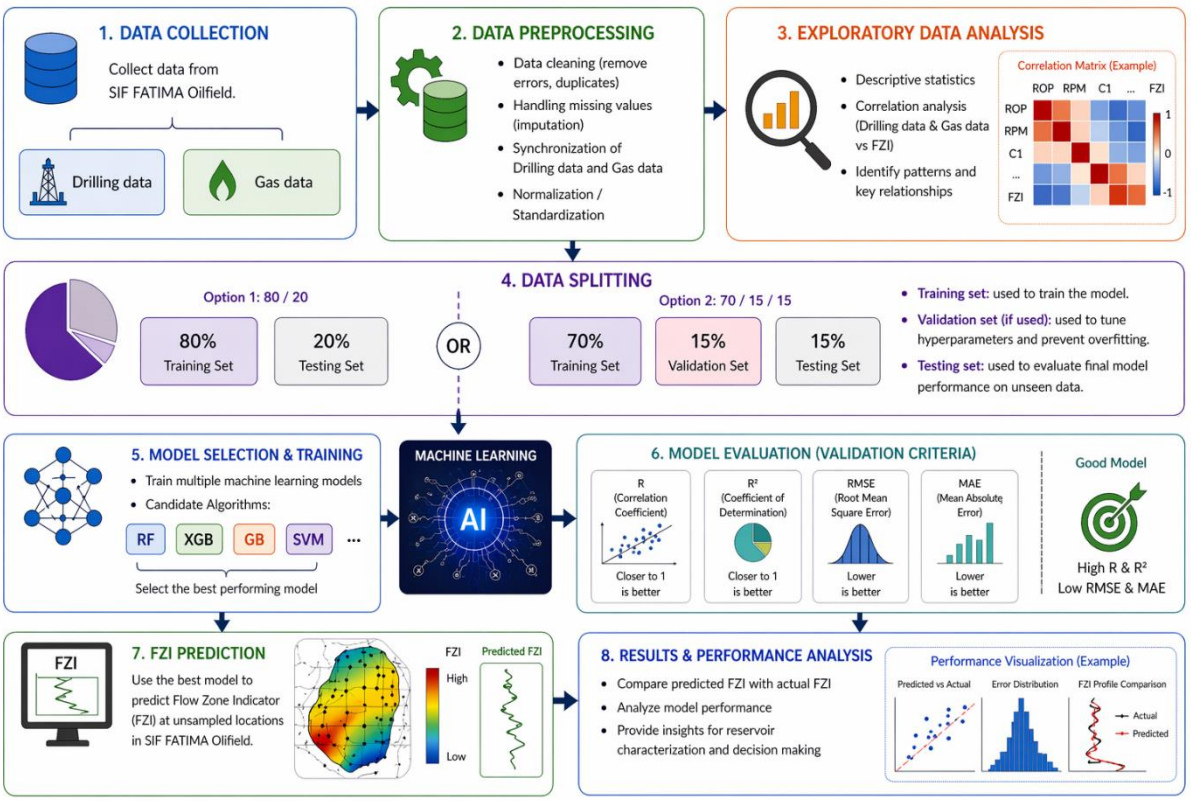


Figure 17. Flowchart summary of Flow Zone Indicator

Conclusions

This workflow made it possible to organize the different steps of FZI prediction in a clear and logical way. Starting from data collection, then data cleaning, statistical analysis, data splitting, algorithm selection, and finally model validation, each stage played an important role in improving the quality of the results. This approach helps to better use drilling and gas data and gives a more reliable estimation of Flow Zone Indicator. In general, it can be considered a useful method for improving reservoir characterization and for better understanding the distribution of flow units in the SIF Fatima field.

Chapter 04
Results & Discussions

Introduction

This chapter is devoted to presenting and discussing the results obtained in this study. After preparing the database, the first step was to examine the drilling parameters through descriptive statistics in order to understand the general behavior of the data and detect any noticeable variation between the variables.

A correlation analysis was then carried out to evaluate the relationships between the input parameters and the Flow Zone Indicator (FZI). This step helps to identify which drilling variables may have a stronger connection with FZI and which variables show weak or limited influence In addition, different combinations of input variables were tested to investigate their effect on the prediction process. This is an important step because the quality of the model does not depend only on the algorithm used, but also on the relevance of the selected input data.

Finally, the performance of the machine learning models is evaluated and compared using statistical criteria. The obtained results provide a clearer idea about the ability of drilling data to predict FZI and support reservoir characterization in a faster and more practical way.

1 Data Preparation

Before applying the machine learning models, the database was first organized in a clear and structured form. The input variables used in this study are drilling parameters, including ROP, WOB, RPM, SPP, Q, Torque, BitRun, and BitTim, while FZI was selected as the output variable.

The data were arranged so that each row represents one observation, and each column represents one parameter. This organization was necessary to make the dataset suitable for the next steps of the study, including statistical analysis, correlation analysis, and machine learning modeling.

In this step, the main objective was to prepare a clean and usable database, where the input parameters and the target variable are clearly defined. After this preparation, the dataset became ready for further analysis and prediction of the Flow Zone Indicator.

2 Results and Discussion

2.1 Univariate Statistics

The table presents the preliminary statistical analysis of the drilling variables and the FZI coefficient; the database contains 196 values for each variable, indicating that there are no missing values at this stage. In general, some variables such as RPM, WOB, and Torque show relatively acceptable variation, while other variables such as ROP and FZI exhibit clear outliers due to the presence of very high values compared to the rest of the values Regarding FZI, we observe that the mean is 7.28, while the median is only ٤2.09, indicating that most values are low to medium, with some high values that raised the mean. This reflects the heterogeneity of the reservoir, as the permeability of the rocks varies from one area to another Overall, this analysis shows that the data is suitable for moving to the modeling phase, but it is necessary to pay attention to outliers, especially in ROP and FZI, as they may affect the performance of machine learning models

Table 2. Univariate Analysis of Input Variables

Statistic	ROP (m/h)	WOB (ton)	RPM (rpm)	SPP (psi)	Q (gpm)	T (kNm)	BitRun (mt)	BitTim (hrs)	FZI (µm)
count	196.00	196.00	196.00	196.00	196.00	196.00	196.00	196.00	196.00
mean	14.01	6.16	110.39	2188.55	173.24	199.88	28.21	4.33	7.29
std	37.54	3.12	17.87	675.48	51.23	48.62	14.87	2.74	17.46
min	2.10	1.51	48.00	1324.00	100.00	109.00	0.45	0.10	0.12
25%	4.57	3.39	103.00	1639.00	118.00	162.25	17.75	2.90	1.33
50%	7.95	4.36	112.50	1728.50	212.00	190.00	29.00	3.70	2.09
75%	15.05	9.09	119.00	2991.00	223.00	223.25	40.00	5.35	3.80
max	517.70	12.01	190.00	3146.00	231.00	344.00	55.00	17.00	118.61

2.2 Bivariate Statistics

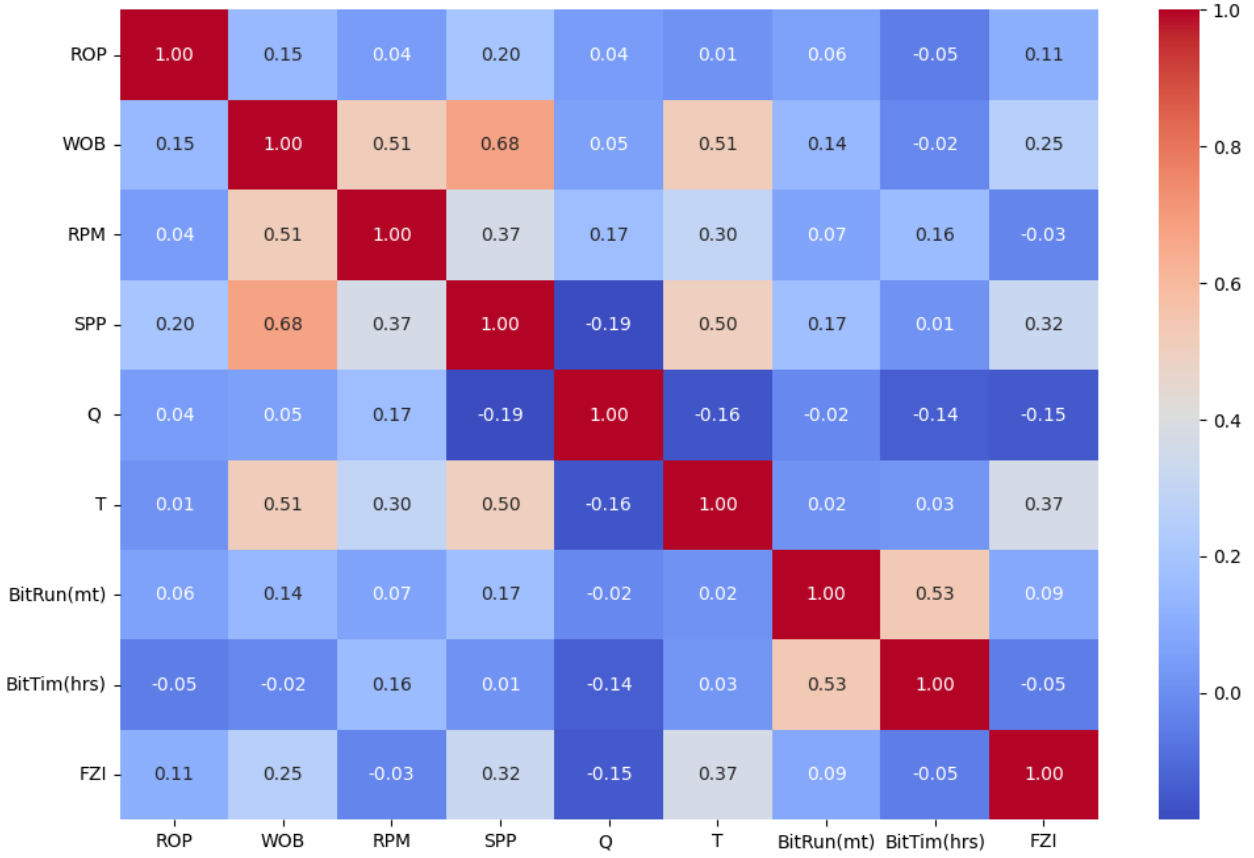


Figure 18. *Bivariate Analysis of Input Variables*

ROP shows a weak correlation with most variables. Its strongest correlation was with SPP at 0.20, followed by WOB at 0.15, and FZI at only 0.11. This means that the penetration rate does not have a clear linear relationship with FZI in this dataset.

- **WOB** has moderate to strong correlations with some drilling variables, particularly with SPP (0.68) and with RPM and Torque (0.51). It also correlates with FZI (0.25), which is a weak to moderate positive relationship. This indicates that the weight of the drill bit may have a partial effect on changes in FZI, but it is not the only factor.
- **RPM** has a moderate correlation with WOB (0.51), with SPP (0.37), and with Torque (0.30). Its correlation with FZI, however, is weak and negative (-0.03), which means that RPM does not appear to have a clear direct effect on FZI.
- **Q** or flow rate, shows weak correlations with most variables, some of which are negative. It is negatively correlated with SPP (–0.19), with Torque (–0.16), and with FZI (–0.15). This indicates that an increase in flow rate in this dataset is not associated with an increase in FZI rather, there is a weak inverse relationship.

- **Torque (T)** shows significant correlations with certain variables, particularly with WOB (0.51) and SPP (0.50). It also has the highest correlation with FZI (0.37) compared to the other variables. This suggests that torque may be one of the most important variables influencing the prediction of FZI.
- **BitTim** has a moderate correlation with BitRun, with a value of 0.53, while its correlation with the other variables is very weak. Its correlation with FZI is negative and weak, with a value of -0.05, indicating that the miller's working hours do not directly explain changes in FZI.
- **BitRun** shows a clear correlation with BitTim, with a value of 0.53, which makes sense because an increase in the distance covered by the runner is often associated with an increase in running time. Its correlation with FZI, however, is very weak, with a value of 0.09, meaning that its direct impact on FZI is limited.
- **FZI** shows its strongest correlation with Torque at 0.37, followed by SPP at 0.32, and then WOB at 0.25. Its relationship with the remaining variables is very weak, such as ROP = 0.11, BitRun = 0.09, RPM = -0.03, BitTim = -0.05, and Q = -0.15.

3 Prediction

3.1 Feature importance and Shap

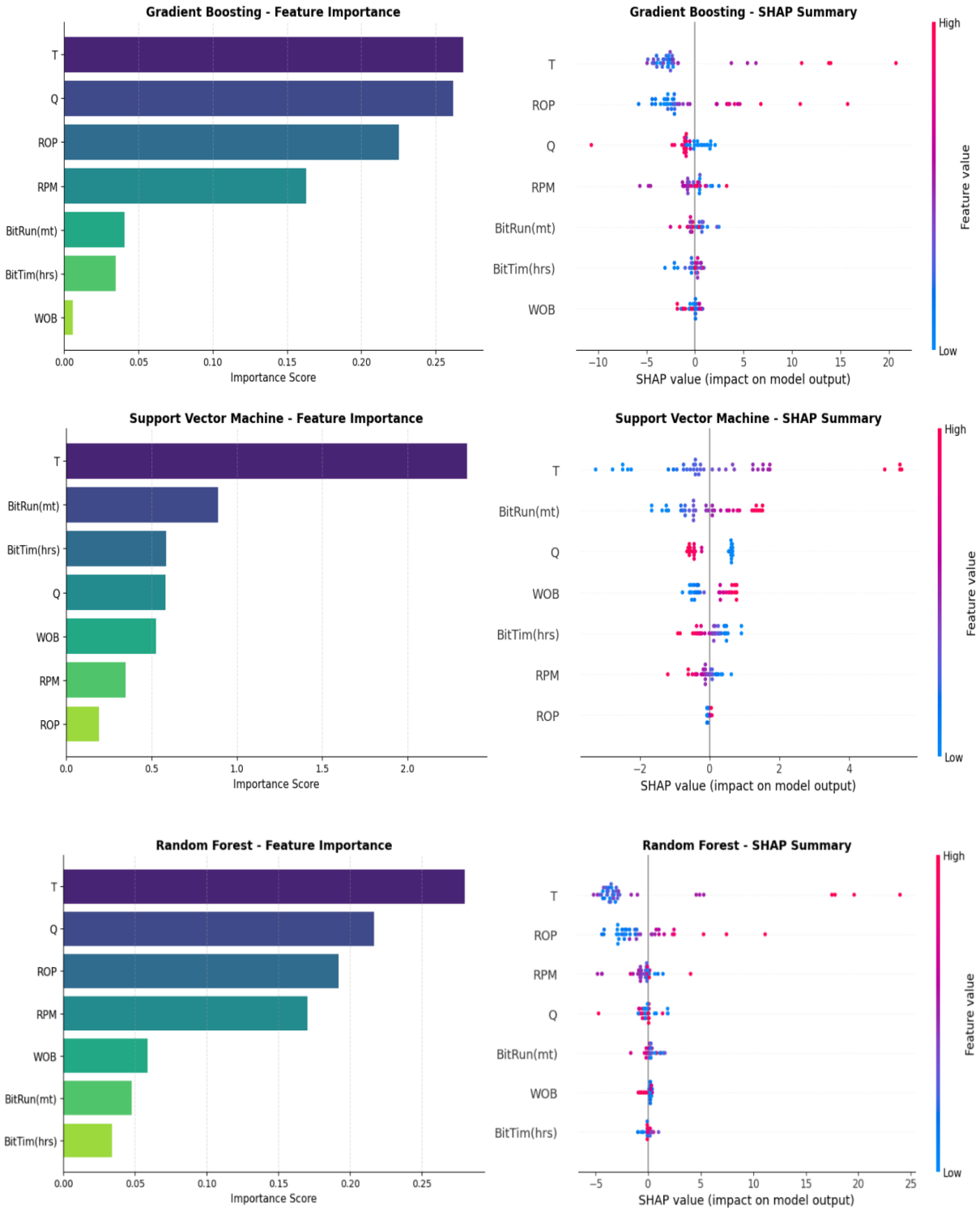


Figure 19. Global Interpretation of FZI Prediction Model Based on Feature Importance and SHAP Summary

3.2 Appalied Machine learning models

Table 3. Prediction phases

Combination	Data Type	Input Variables	Output
1	DD	ROP, WOB, RPM, Q, Torque, BitTim, BitRun	**FZI**
2	DD	ROP, WOB, RPM, Q, Torque, BitRun	**FZI**
3	DD	ROP, WOB, RPM, Q, Torque	**FZI**
4	DD	ROP, RPM, Q, Torque	**FZI**

In this step, different sets of drilling parameters were tested to assess their impact on predicting the Flow Zone Index (FZI). The goal was to determine whether all input variables were necessary, or whether fewer parameters were sufficient to achieve good prediction results.

Four input sets were prepared. The first set includes all available drilling parameters: rate of penetration (ROP), weight on bit (WOB), rotational speed (RPM), flow rate (Q), torque, drilling time (BitTim), and bit run time (BitRun). In the following sets, some variables were gradually removed

This step is important because the accuracy of machine learning models depends not only on the algorithm used, but also on the quality and relevance of the selected input variables. Therefore, comparing these sets helps identify the most useful drilling parameters for predicting the Flow Zone Index (FZI).

3.3 Performance evaluation

Table 4.ML model results

ML Models	Training Set		Testing Set	
	R^2	RMSE	R^2	RMSE
Drilling Data				
RF 1	0.882146	6.520604	0.655554	5.100388
RF 2	0.893259	6.205561	0.684689	4.879915
RF 3	0.872267	6.788396	0.739177	4.438285
RF 4	**0.880700**	**6.560481**	**0.809884**	**3.789239**
GB 1	**0.998840**	**0.646776**	**0.805348**	**3.834171**
GB 2	0.998794	0.659545	0.737168	4.455343
GB 3	0.997986	0.852306	0.731069	4.506744
GB 4	0.998047	0.839291	0.800133	3.885194
SVM 1	0.042038	18.590443	0.254786	7.502100
SVM 2	**0.060691**	**18.408568**	**0.302829**	**7.256248**
SVM 3	0.069542	18.321626	0.287018	7.338068
SVM 4	0.089134	18.127713	0.253522	7.508459

3.4 Cross-Plots of ML Models by Scenario

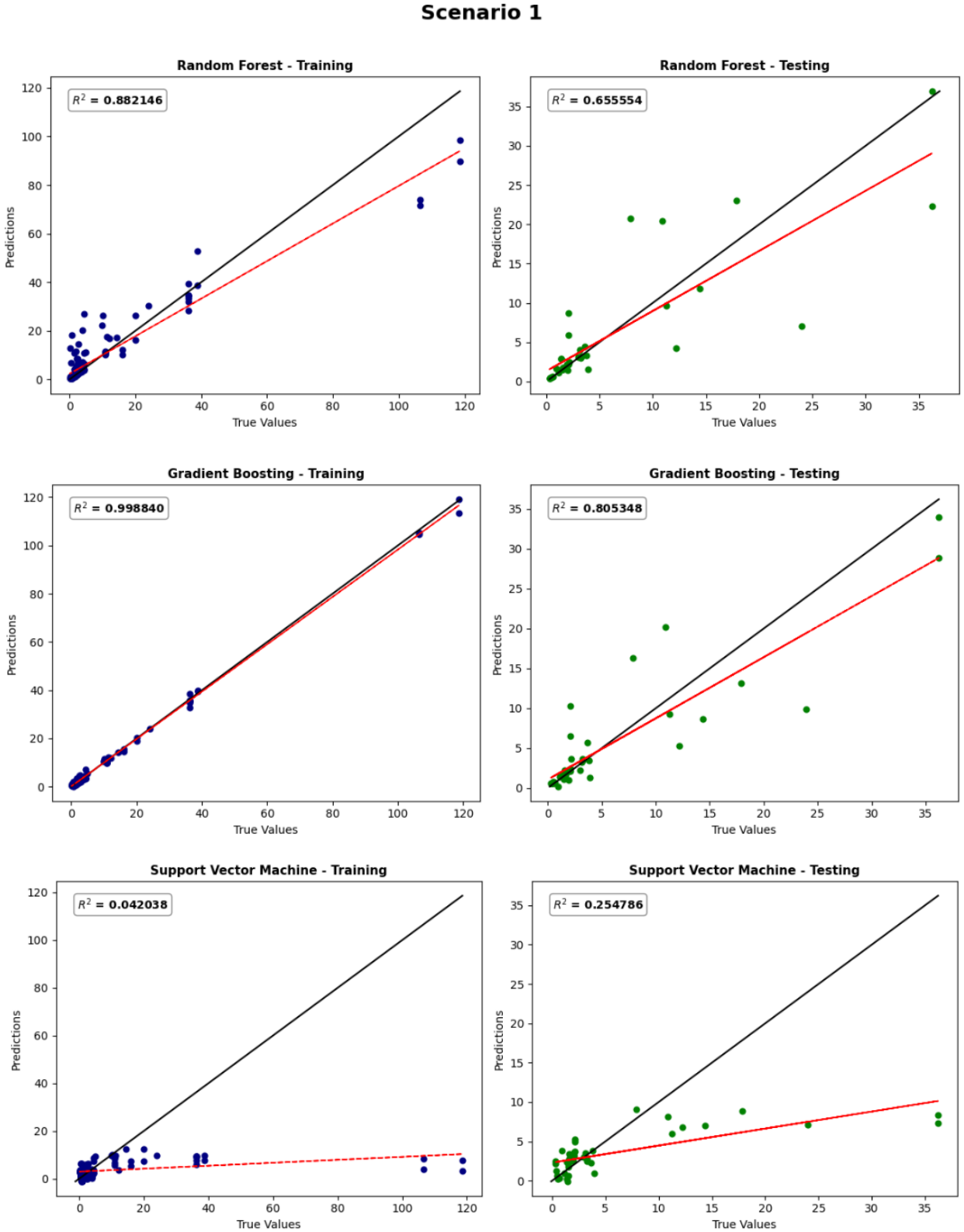


Figure 20. Cross Plot Analysis for Assessing FZI Prediction Accuracy

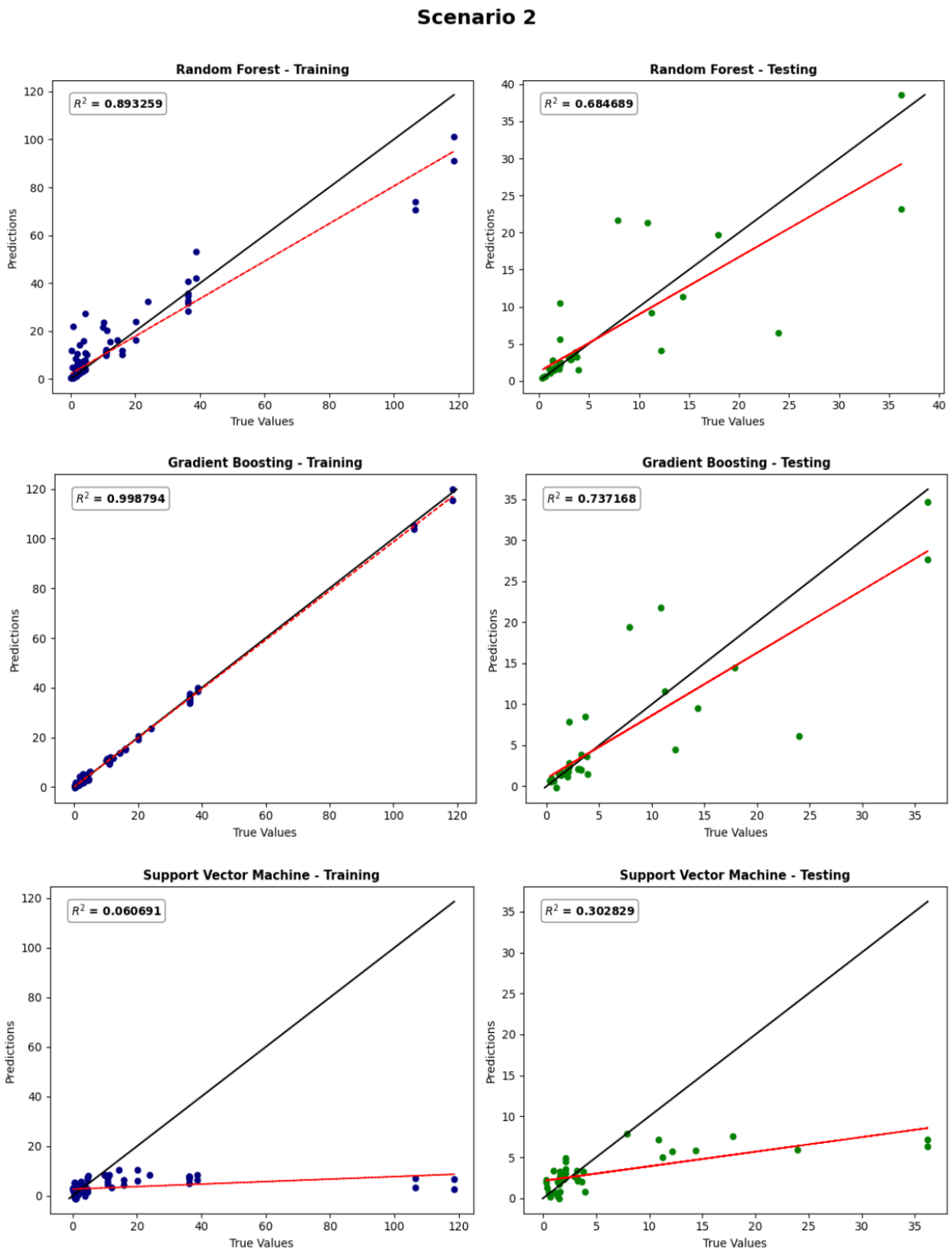


Figure 21. Cross Plot Analysis for Assessing FZI Prediction Accuracy

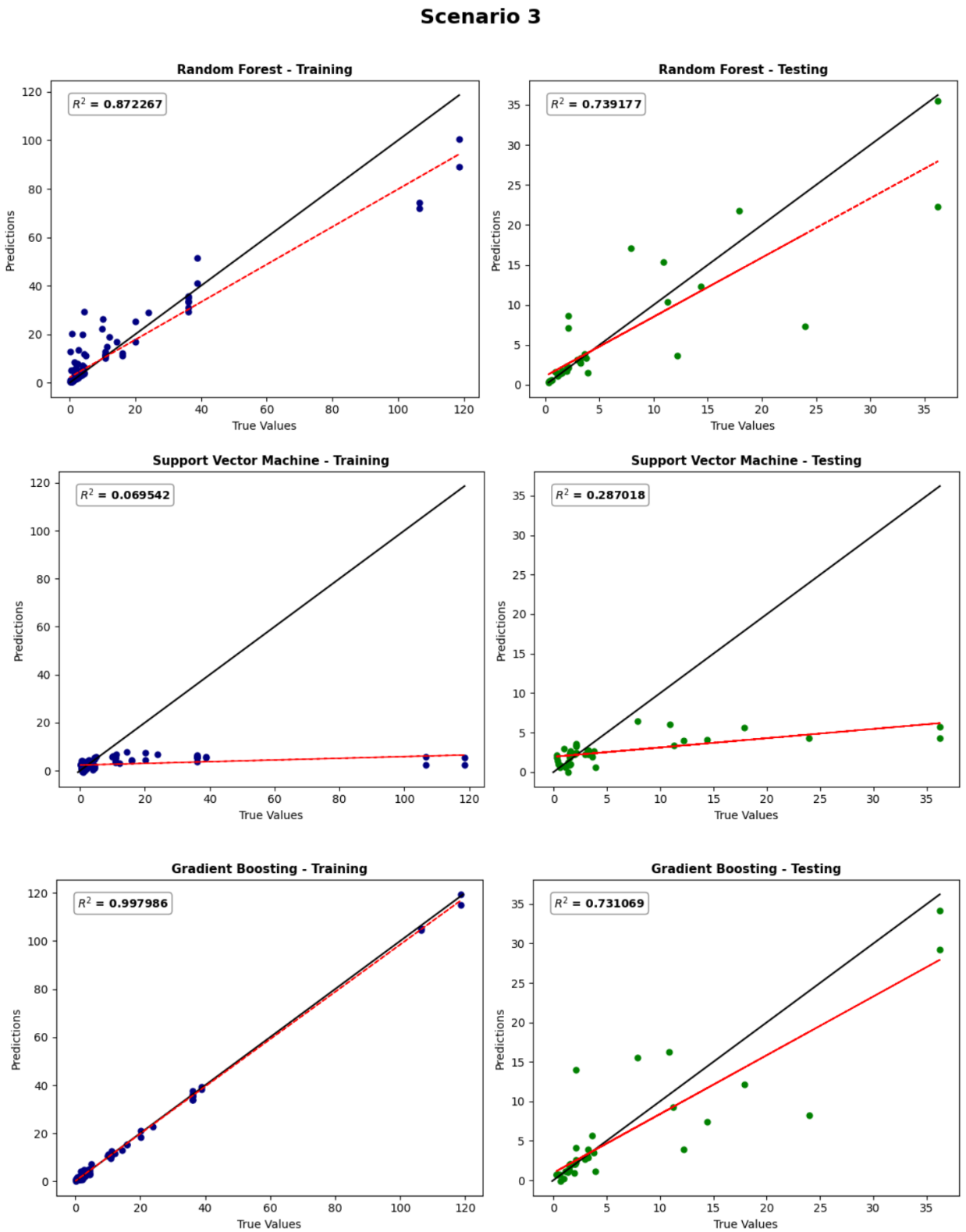


Figure 22. Cross Plot Analysis for Assessing FZI Prediction Accuracy

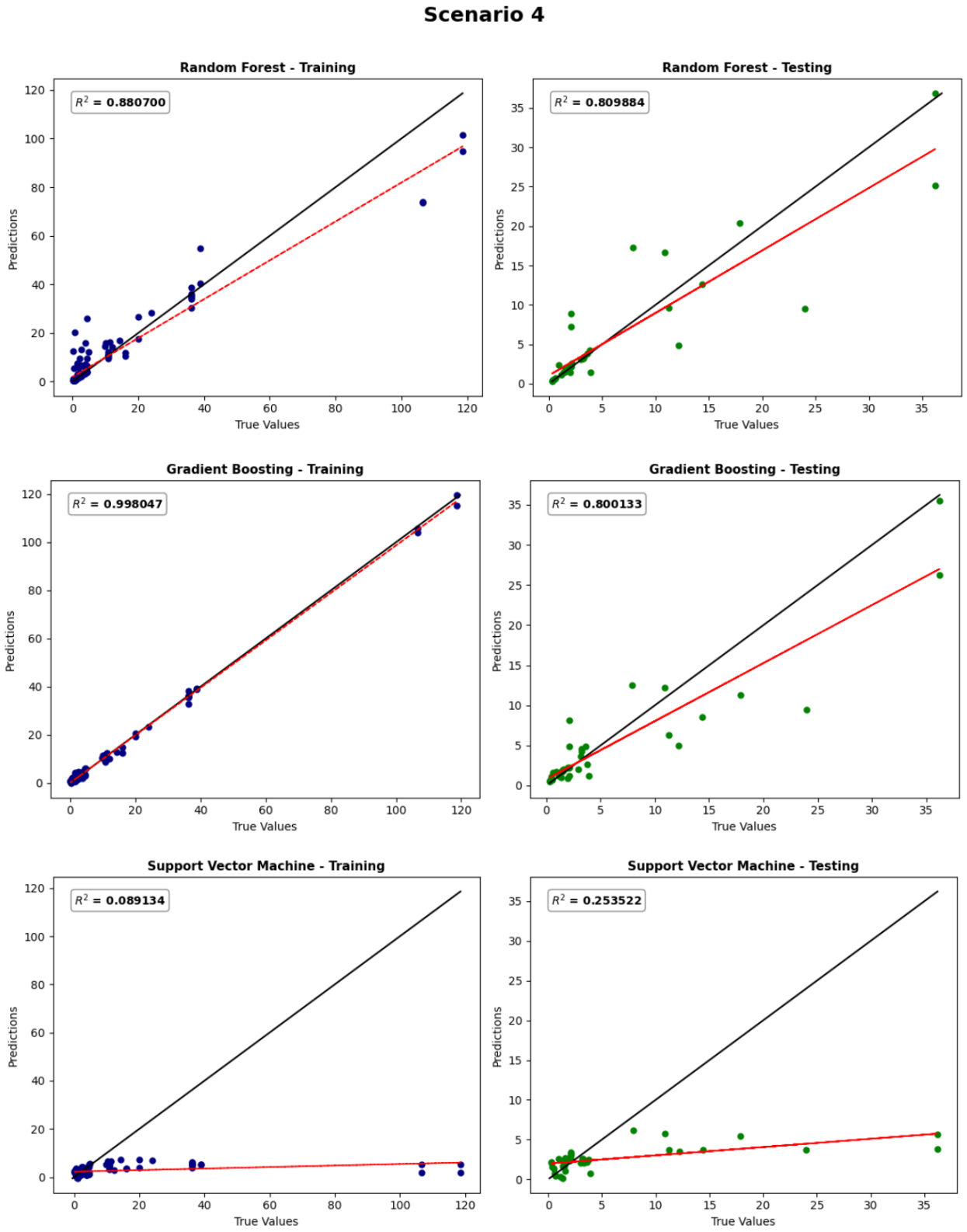


Figure 23. *Cross Plot Analysis for Assessing FZI Prediction Accuracy*

4 Performance Evaluation of Machine Learning Algorithms

The performance of three machine learning models Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) was evaluated for the prediction of the Flow Zone Index (FZI) based on surface drilling parameters (ROP, WOB, RPM, Q, T, BitTim, and BitRun). The results demonstrated a clear variation in each model's ability to capture the complex, non-linear relationships linking drilling dynamics to petrophysical properties.

4.1 Gradient Boosting (GB) Model

The Gradient Boosting model proved to be the most accurate and reliable in this study. Leveraging its sequential error-correction mechanism, the model achieved near-optimal performance during training (R^2= 0.998) while maintaining excellent generalization capability on unseen data during testing (R^2= 0.805). This indicates that the Gradient Boosting algorithm is highly effective at capturing intricate interactions between parameters, such as flow rate (Q) and rate of penetration (ROP), without significant overfitting.

4.2 Random Forest (RF) Model

The Random Forest model demonstrated stable and robust performance, recording R^2 values of (0.882) for training and (0.655) for testing. Although the bagging approach effectively reduced variance, the drop in predictive accuracy during the testing phase compared to the GB model suggests that the model experienced a degree of overfitting, or encountered limitations in mapping the deeper non-linear patterns of drilling data associated with the FZI, as compared to boosting algorithms.

4.3 Support Vector Machine (SVM) Model

The SVM model exhibited the weakest performance in this comparison, with R^2 values of (0.042) for training and (0.254) for testing. This underperformance is primarily attributed to the use of a linear kernel. Given that the relationships between surface drilling parameters and reservoir properties (such as FZI) are characterized by complex, non-linear geological and physical interactions, a linear separator proves insufficient for accurate modeling, rendering this configuration unsuitable for the current application without significant adjustments to its parameters.

GENERAL CONCLUSION

GENERAL CONCLUSION

This research work was conducted within the framework of improving reservoir characterization through the prediction of the Flow Zone Indicator (FZI) using mud logging data and explainable machine learning techniques. In heterogeneous reservoirs, such as the TAGI reservoir of the Sif Fatima field, the evaluation of reservoir quality is often difficult because petrophysical properties such as porosity, permeability, and pore connectivity can vary significantly from one interval to another. For this reason, the integration of artificial intelligence methods represents a promising solution for estimating reservoir quality in a faster, more continuous, and more cost-effective way.

The study first presented the theoretical background of artificial intelligence, machine learning, and deep learning, with particular focus on their applications in petroleum engineering and reservoir characterization. It also discussed the geological and reservoir context of the Berkine Basin and the Sif Fatima field. The geological analysis showed that the TAGI reservoir is mainly controlled by fluvial facies, where clean sandstone channels generally represent good reservoir zones, while clay-rich intervals reduce permeability and create barriers to fluid flow.

In the methodological part, both conventional and unconventional approaches were discussed. The conventional calculation of FZI depends mainly on core-derived porosity and permeability, which are accurate but expensive, discontinuous, and limited to selected intervals. To overcome these limitations, this study proposed the use of drilling parameters and mud logging data as input variables for machine learning models. The workflow included data collection, data cleaning, descriptive statistical analysis, data splitting, algorithm selection, and model validation using performance criteria such as R^2 and RMSE.

The results showed that drilling parameters can provide useful information for predicting FZI. The correlation analysis indicated that Torque, SPP, and WOB have the strongest relationships with FZI compared with the other variables. Several machine learning models were tested, including Random Forest, Gradient Boosting, and Support Vector Machine. Among these models, Random Forest and Gradient Boosting gave the best predictive performance, while SVM showed weaker results. The best performance was obtained with the reduced input combination in Scenario 4, where Random Forest achieved the highest testing accuracy and the lowest prediction error.

Moreover, the results confirm that the prediction of FZI using machine learning can help identify productive intervals, classify hydraulic flow units, and improve the understanding of reservoir heterogeneity. This approach reduces dependence on costly core data and allows better use of real-time drilling information. It can therefore support faster operational decisions during drilling and contribute to better reservoir management in the Sif Fatima field.

Recommendations

Looking forward, this research opens several opportunities for further improvement. First, it is recommended to increase the size and diversity of the dataset by including more wells from the Sif Fatima field and nearby fields in the Berkine Basin. A larger database would improve the robustness and generalization capacity of the machine learning models.

Second, future studies should include additional data types such as well logs, core analysis, seismic attributes, and production data. The integration of these complementary datasets could improve the accuracy of FZI prediction and provide a more complete description of reservoir quality.

Third, more advanced machine learning and deep learning models could be tested, such as Artificial Neural Networks, XGBoost, hybrid models, or optimized ensemble methods. Hyperparameter optimization techniques may also improve model performance and reduce prediction errors.

Finally, explainable artificial intelligence methods, especially SHAP analysis, should be further developed to better understand the contribution of each input parameter to the predicted FZI values. This would make the model more transparent and increase confidence in its use as a decision-support tool in petroleum reservoir studies.

In conclusion, this study demonstrates the feasibility and usefulness of applying machine learning to predict the Flow Zone Indicator from mud logging and drilling data. The proposed approach represents a practical and data-driven solution for reservoir characterization, especially in heterogeneous and partially cored formations. By combining geological understanding, drilling data, and artificial intelligence, this work

contributes to more efficient, cost-effective, and intelligent decision-making in petroleum exploration and reservoir development.

BIBLIOGRAPHY

AbdelJabbar, A. A., et al. (2020). Prediction of porosity and permeability from drilling data using machine learning techniques.

Abnavi, A. D., Torghabeh, A. K., & Qajar, J. (2021). Hydraulic flow units and ANFIS methods to predict permeability in heterogeneous carbonate reservoirs: Middle East gas reservoir. Arabian Journal of Geosciences, 14, 754.

Adnan, M., & Abed, A. (2014). Hydraulic flow units and permeability prediction in a carbonate reservoir, southern Iraq from well log data using non-parametric correlation. Science Technology Engineering, 3(1), 480–486.

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). MIT Press.

Al-Sabaa, M., et al. (2021). Real-time reservoir property estimation using machine learning methods.

Amaefule, J. O., Altunbay, M., Tiab, D., Kersey, D. G., & Keelan, D. K. (1993). Enhanced reservoir description: Using core and log data to identify hydraulic flow units and predict permeability in uncored intervals/wells. SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Ameur-Zaimeche, O. (2014). Modélisation et reconstitution des faciès non carottés à l'aide des méthodes statistiques multivariées du réservoir TAGI : application au champ de Sif Fatima, bassin de Berkine (Mémoire de magister). Université Kasdi Merbah, Ouargla.

Ameur-Zaimeche, O., et al. (2022). Porosity prediction using mud gas and drilling data through machine learning algorithms.

Aminzadeh, F., & de Groot, P. (2006). Neural networks and their applications in the oil and gas industry. EAGE Publications.

Aminzadeh, F., Temizel, C., & Hajizadeh, Y. (2022). Applications in reservoir characterization and field development optimization. In Artificial Intelligence and Data Analytics for Energy Exploration and Production (pp. 271–311). Wiley.

Anifowose, F. A. (2011). Artificial intelligence application in reservoir characterization and modeling: Whitening the black box. SPE Saudi Arabia Section Young Professionals Technical Symposium. https://doi.org/10.2118/155413-MS

Anifowose, F. A., et al. (2024). Porosity prediction from mud gas datasets using machine learning approaches.

Arigbe, O., et al. (2019). Neural-network-based permeability prediction from drilling data.

Azoug, Y., et al. (2007). Well Evaluation Conference (WEC Algeria 2007). Sonatrach–Schlumberger.

Bajaj, S. (2021). Artificial intelligence in upstream oil and gas: Top use cases and benefits. Birlasoft.

Bengio, Y., Goodfellow, I., & Courville, A. (2016). Deep learning. MIT Press.

Binetti, M., et al. (2024). Machine learning integration with remote sensing and environmental datasets.

Boumaza, M., & Semai, F. (2014). Évaluation quantitative et qualitative d'un réservoir : Cas du réservoir TAGI–SIF Fatima, bassin de Berkine (Algérie orientale).

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Carman, P. C. (1997). Fluid flow through granular beds. Chemical Engineering Research and Design, 75, S32–S48.

Chen, Y., et al. (2020). MLSTM-based prediction of porosity and permeability using drilling information.

Dahoumane, O., & Guerfi, L. (2024). Caractérisation géochimique des niveaux roche mère du Frasnien et Fammenien-Strunien et modélisation 1D de la région de Rhourde El Fares (bassin de Berkine). Université Mouloud Mammeri.

Elmorsy, M., et al. (2022). Permeability prediction using ANN and Random Forest methods.

Ezekiel, et al. (2022). Machine-learning-based permeability prediction in heterogeneous reservoirs.

Galeazzi, S., et al. (2010). Regional geology and petroleum systems of the Illizi–Berkine area of the Algerian Saharan Platform: An overview. Marine and Petroleum Geology, 27(1), 143–178.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hassaan, et al. (2024). Real-time porosity and permeability estimation from drilling data using explainable machine learning.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260.

Kelai, A. (2025). Real-time prediction of formation permeability using drilling data and explainable machine learning: A case study from the Hassi Tarfa Field.

King Fahd University of Petroleum & Minerals. (2021). Application of artificial intelligence in petroleum industry.

Kozeny, J. (1927). Über kapillare Leitung des Wassers im Boden. Sitzungsberichte der Akademie der Wissenschaften in Wien, 136, 271–306.

Lawal, K. A., Alhassan, M., & Salami, A. (2024). Machine learning applications in reservoir characterization and petrophysical property prediction. Petroleum Research, 9(1), 45–61.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.

Mehrabi, H., Kadkhodaie-Ilkhchi, A., & Rahimpour-Bonab, H. (2020). Prediction of flow zone indicator using well log data and machine learning methods. Journal of Petroleum Exploration and Production Technology, 10(4), 1567–1581.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Mohaghegh, S. (2020). Data-driven reservoir management using artificial intelligence. SPE.

Mohammadian, et al. (2022). XGBoost-based porosity estimation from drilling and logging data.

Nande, et al. (2022). Machine learning approaches for permeability prediction.

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons.B, 4, 51–62.

Ouladmansour, et al. (2023). Porosity prediction from drilling data using machine learning techniques.

Otchere, D., et al. (2021). Water saturation prediction using machine learning models.

Omiri, M., & Dahmani, A. (2021). Découpage séquentiel, évaluation pétrophysique des réservoirs SAG et TAG du bassin de Berkine. Université Mouloud Mammeri.

Peng, et al. (2023). Machine learning for environmental systems analysis and prediction.

Qalandari, et al. (2022). ANN and XGBoost models for permeability prediction from integrated datasets.

Raza, M., Ali, R., & Al-Anazi, A. (2020). Applications of artificial intelligence in petroleum industry. Journal of Petroleum Science and Engineering, 193, 107418.

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

Sadeghpour, M., Rahimpour-Bonab, H., & Tokhmechi, B. (2025). Explainable machine learning approaches for reservoir quality and flow zone indicator prediction using drilling and petrophysical data. Journal of Petroleum Science and Engineering, 245, 111245.

Salcedo-Sanz, S., et al. (2020). Machine learning in Earth sciences and remote sensing applications.

Santhanam, et al. (2016). Extreme gradient boosting for predictive analytics.

Sediri, F. (2017). Étude des unités du Silurien argileux gréseux de l’ouest du bassin de Berkine. Université Mouloud Mammeri.

Sharma, et al. (2023). Porosity and density prediction using gamma ray and drilling data.

Tiab, D., & Donaldson, E. C. (2022). Petrophysics: Theory and practice of measuring reservoir rock and fluid transport properties (4th ed.). Gulf Professional Publishing.

Tian, Y., et al. (2021). Real-time reservoir property estimation using machine learning and MWD data.

Tembely, M., et al. (2021). Reservoir characterization using machine learning techniques.

WEC (Well Evaluation Conference). (2007). Rapport inédit. Sonatrach

Wang, X., et al. (2020). Porosity and permeability prediction using machine learning models.

Zhu, D., Hill, A. D., et al. (2018). Modern reservoir characterization techniques using integrated petrophysical and geological data. SPE Reservoir Evaluation & Engineering, 21(3), 455–468.

Zolotukhin, A., et al. (2019). Neural-network-based permeability prediction using drilling parameters.

Zou, J., Han, Y., & So, S.-S. (2009). Overview of artificial neural networks. In Artificial Neural Networks: Methods and Applications (pp. 14–22).